AD________________

Award Number: DAMD17-02-1-0364

TITLE: Genome-Wide Nucleic Acid/Protein Interaction in Breast Cancer

PRINCIPAL INVESTIGATOR: Pamela A. Silver, Ph.D.

CONTRACTING ORGANIZATION: Dana-Farber Cancer Institute
Boston, Massachusetts 02115

REPORT DATE: April 2005

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
April 2005

3. REPORT TYPE AND DATES COVERED
Final (15 Apr 2002 - 14 Mar 2005)

4. TITLE AND SUBTITLE
Genome-Wide Nucleic Acid/Protein Interaction in Breast Cancer

5. FUNDING NUMBERS
DAMD17-02-1-0364

6. AUTHOR(S)
Pamela A. Silver, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Dana-Farber Cancer Institute
Boston, Massachusetts 02115

E-Mail: pamela_silver@dfci.harvard.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)
Since many types of breast cancer remain untreatable, the research proposal aims to develop novel genomic technology to identify potential therapeutic targets and to aid in diagnosing various types of breast cancer at the molecular level. The overarching goal of the proposal is to develop a technology to screen nucleic-acid protein interactions on a genome scale with a focus on understanding complexes involved in breast cancer. In order to identify the regulatory networks of interactions between RNAs and proteins, we proposed to develop a rapid genome-scale method to determine the specific RNA targets and RNA binding sites of proteins. The aims were to 1) discover RNA targets of specific RNA binding proteins and 2) define the RNA sequences recognized by proteins using novel nanotechnologies including development of optically encoded beads containing both a unique optical signature and a specific oligonucleotide. This technology is being complemented by genome-wide chromatin immunoprecipitation and RNA profiling on exon arrays. Progress during the past year has been made on Aims 2, 3 and 4 of the original grant as detailed below. Two papers mapping proteins to the genome have been submitted for publication while a third paper is in preparation.

14. SUBJECT TERMS
RNA-protein interactions, DNA-protein interactions, genomic screening, regulation of gene expression, technology development

15. NUMBER OF PAGES
86

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
Unlimited

Table of Contents

INTRODUCTION:

Since many types of breast cancer remain untreatable, the research proposal aims to develop novel genomic technology to identify potential therapeutic targets and to aid in diagnosing various types of breast cancer at the molecular level. The overarching goal of the proposal is to develop a technology to screen nucleic-acid protein interactions on a genome scale with a focus on understanding complexes involved in breast cancer. In order to identify the regulatory networks of interactions between RNAs and proteins, we proposed to develop a rapid genome-scale method to determine the specific RNA targets and RNA binding sites of proteins. The aims were to 1) discover RNA targets of specific RNA binding proteins and 2) define the RNA sequences recognized by proteins using novel nanotechnologies including development of optically encoded beads containing both a unique optical signature and a specific oligonucleotide. This technology is being complemented by genome-wide chromatin immunoprecipitation and RNA profiling on exon arrays. Progress during the past year has been made on Aims 2, 3 and 4 of the original grant as detailed below. Two papers mapping proteins to the genome have been submitted for publication while a third paper is in preparation.

BODY:

Aim 1 was successfully completed as marked by the publication of a paper describing the assay (Brodsky and Silver, 2002). Thus, we have laid the groundwork for genomic and small molecule screening using the microbead assay. The assay has generated some interest in the community as we have recently written an invited review discussing the microbead assay we have developed [1] (see attached). Although the goals remain the same, we are now using a new and potentially more powerful method than originally proposed in Aim 2. The research accomplishments associated with each task outlined in the approved Statement of Work are detailed below.

Technical Objective 1: Determine optimal conditions for bead-based genomic screening.

The goals of this objective were completed as outlined in a published paper [1] (see attached). We were also invited to write a review highlighting this new technology to probe RNA-protein interactions [2] (see attached).

Technical Objective 2: Identification of target RNAs of clinically important proteins.

While the new microbead technology is being developed, current microarray technologies could be used to determine candidate binding targets. Recently, Chromatin Immunoprecipitation (CHiP) has emerged as a powerful method to identify where on a gene and - in combination with microarrays (ChiP-chip) - on which genes chromatin associated proteins are binding [3, 4]. Briefly, cells are cross-linked and chromatin is sheared to approximately 1,000 bp average size. The protein of interest is immunoprecpitated and the DNA is isolated for quantitative PCR analysis or microarrays. Because our lab, as well as others, has shown that many RNA binding proteins bind co-transcriptionally, we can take advantage of this approach. In addition, representing a more *in vivo* situation, a potentially significant advantage of CHiP is that the cells are formaldehyde cross-linked allowing the capture of dynamic interactions. We have modified and improved the ChIP approach to localize RNA Binding Proteins (RBPs) on various genes. Some of our modifications include the use of a second protein-protein crosslinker in addition to the commonly used formaldehyde. We also use log-linear fitting of real-time PCR data which enhances our sensitivity and dynamic range of the analysis [5]. Figure 1 shows enrichments of two RNA binding proteins and two states of RNA Polymerase II across the PTB gene. We observe RNA binding proteins at the 5' end of genes as well as at sites of alternative splicing. Similarly, we observe hypophosphorylated RNA polymerase II at the 5' end of the PTB gene but not at other locations of the PTB gene.

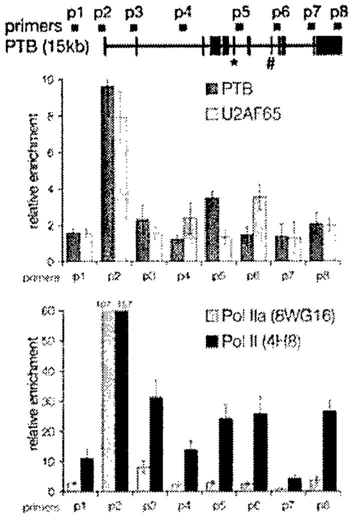
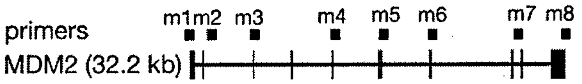


Figure 1. ChIP assay reveals patterns of RNA binding protein and RNA Polymerase II enrichment across genes. A series of primers A. The two RNA binding proteins PTB and U2AF65 are found at the 5' end and at alternatively spliced exons. Hypophosphorylated RNA Polymerase II (PolIIa) is found only around the transcription initiation sites. Phosphorylated RNA Polymerase II is found at the transcription initiation sites, the alternatively spliced exons, as well as the 3' end of the gene. These latter locations are regions where transcription is coupled to pre-mRNA processing.



	m1	m2	m3	m4	m5	m6	m7	m8
PTB	+	+	+	+	-	+	+	+
U2AF65	+	+	-	+	+	+	-	-
Pol IIa	+	+	-	-	-	-	-	-
Pol II	+	+	-	-	-	+	+	+
PTB RNAi	▬	▬	▬		↓	↓		
U2AF65 RNAi	↓	↓	↓		↑	↑		

Figure 2. Summary of factor enrichments across the MDM2 gene. + indicates enrichment and – indicates no significant enrichment at each location. Upon siRNA knockdown of PTB or U2AF65, levels of specific combinations of exons were monitored by RNAse protection assays.

Importantly, we have shown that the sites of enrichment are regulated by PTB and U2AF65. Figure 2 shows the enrichment of PTB and U2AF65 across the MDM2 oncogene. MDM2 is an important oncogene mutated in many tumors and tumor specific isoforms have been identified [6]. Knockdown experiments using siRNAs against PTB and U2AF65 suggest that the exon levels of some of these enriched regions are regulated by PTB and U2AF65 as determined by RNAse Protection Assays (RPA). Interestingly, the region around the m5 and m6 primers is in the vicinity of the nuclear localization sequence suggesting that PTB and U2AF65 may be regulating the inclusion or exclusion of these protein sequences. Thus, these RNA binding proteins may be controlling the cytoplasmic localization of MDM2. These studies are providing insight into some of the factors involved in the complicated post-transcriptional regulation of MDM2.

We have also found that PTB and U2AF65 are recruited to a wide variety of genes at the 5' end. We observe PTB and U2AF65 at the 5' end of intronless genes such as histones, constitutively spliced genes such as actin, and alternatively spliced genes. In Technical Objectives 3 and 4, we describe microarray data across hundreds of genes where we observe the generality of these observations.

In sum, these data demonstrate that ChIP of RNA binding proteins in mammalian cells is feasible. Furthermore, we find that the sites of recruitment are locations of regulation of gene expression.

Technical Objective 3: Screen RNA-protein interactions of a targeted set of genes.

To determine the binding profile of a larger number of genes we designed and constructed our own microarrays to probe a larger number of genes. Open Reading Frames (ORFs) were cloned from cDNA libraries, sequence verified and spotted onto slides. This array includes a number of genes relevant to breast cancer including MDM2 and numerous kinases. ChIP-chip experiments with these ORF microarrays identified a number of new putative targets for PTB including SNK, DAPK3 and MDM2. These genes are rich in alternative splicing. We have verified DAPK3 and MDM2 by PCR analysis. Interestingly, DAPK3 shows approximately 10-fold stronger enrichment at its 3' end than we have observed for any other region for PTB.

With the rapid advances in microarray technology over the past couple of years, we are now reaching genome-scales. We are part of an early access program to use Affymetrix tiled arrays. These arrays include a 25mer oligonucleotide probe every 20 base pairs across the ENCODE regions [7].

In order to learn how to perform ChIP-chip experiments and develop analysis tools, we first analyzed the localization of two states of RNA Polymerase II. These data have been submitted for publication and the manuscript is included in the appendix.

We made a number of technical advances during this work which helped us improve the quality of the data. These include a new random primer amplification method. We have also developed a number of analysis tools which we expect to make generally available including the analysis of constitutive and alternatively spliced exons. This analysis allowed us to discover that hyperphosphorylated RNA Polymerase II accumulated more often at alternatively spliced exons (see attached manuscript).

We are analyzing ENCODE ChIP-chip data for a number of RNA binding proteins with a variety of putative functions. Herein, we will highlight data for two: PTB and Aly. Aly is also known as REF1, RNA Export Factor 1, and is a putative RNA export factor and part of the exon junction complex.

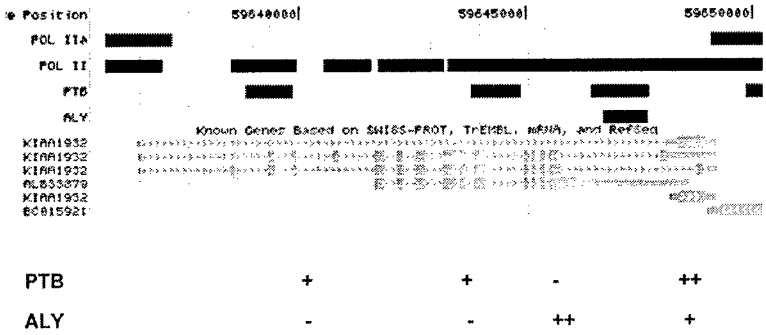
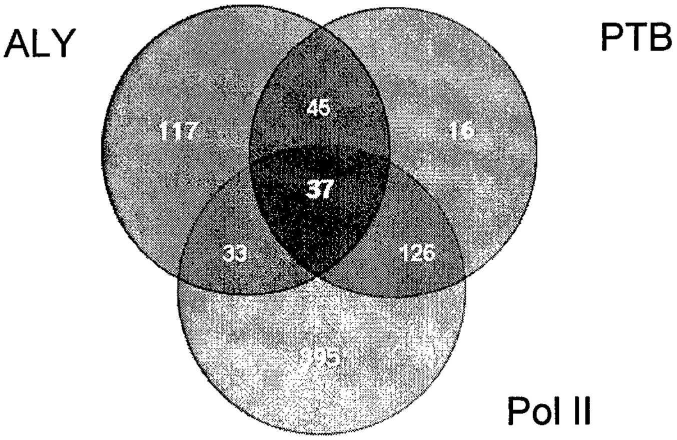


Figure 3. RNA Polymerase II, PTB and ALY localization across the KIAA1932 gene. The black bars represent regions of significant enrichment. The different known Genes transcripts are in blue. PCR data are summarized in red with plus signs indicating significant enrichment. The PCR and array data correlate. Note that PTB localizes to alternative exons while Aly appears to be biased towards the 3' end of the gene.

Figure 4. Venn diagram showing the overlap between ALY, PTB and Pol II. 73% of PTB sites overlap with Pol II while only 30% of ALY sites overlap with PolII. Thus, many ALY sites are not associated with RNA Polymerase II transcription.



An emerging theme is a growing family of dual activity proteins that bind RNA and also regulate transcription. Furthermore, some RNA binding proteins, such as hnRNPK, have been associated with transcription control [8]. We have mapped PTB and Aly to certain promoter regions but not others. Thus, for the first time, we can design experiments to probe the role of these proteins in transcription control using a luciferase reporter assay.

In sum, these data are not only identifying known genes that these RNA binding proteins may be regulating but also providing new insights into how RNA binding proteins may be interacting with the genome. For example, many Aly sites are not in annotated gene regions. One hypothesis is that these sites may be involved in transcription with RNA Polymerase I or III. We are currently exploring what the function of these sites may be.

Our interest in RNA binding proteins and their potential role in breast cancer lead us to collaborate with Myles Brown's group to explore how Estrogen Receptor alpha (ER) is interacting with the genome. Our overlapping interests in developing ChIP-chip technology lead us to work together to map ER across chromosomes chr21/22. A paper has been submitted for publication and is included in the appendix.

Technical Objective 4. Analyze RNA-protein interactions on a genomic scale.

Whole genome Affymetrix tiled arrays have become available to us in late 2004. As a first step towards developing analysis tools to handle this very large scale of data, we localized estrogen receptor and RNA Polymerase II in breast cancer epithelial cell line (MCF-7). Figure 5 shows that we identify thousands of ER and PolII sites across the human genome. Similar, summarizes the number of sites and what kind of sites we observe.

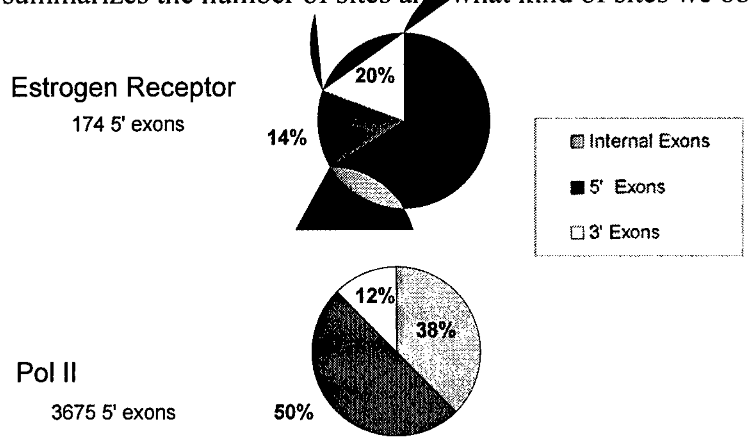


Figure 5. Pie charts show the distribution of the ER and Pol II on different types of exons across the all the non-repetitive DNA in the human genome. Few ER sites are in annotated promoter regions while many of the Pol II sites are.

Pol II on tiled arrays in estrogen treated breast cancer cells gives a list of the genes being actively transcribed as well as novel intergenic regions that may be expressed. Interestingly, we identify 5 microRNAs with significant ER enrichment nearby. Only two of these have both Pol II and ER significant enrichment: mir-223 and mir-152. Current experiments are exploring whether the levels of these and other microRNAs are regulated in an estrogen dependent manner.

To complement these genomic location studies, we are profiling potential splicing changes in response to estrogen using Affymetrix exon arrays as part of a collaboration with Affymetrix. These arrays include probes targeting every known exon in the human genome. Our understanding of cell's response to estrogen remains mysterious despite significant effort exploring transcriptional control using standard gene expression microarrays. Early efforts exploring alternative splicing response suggest that the genes changing at their exon usage are not the same genes whose mRNA levels are changing. Thus, an important part of defining a cell's expression program is alternative splicing. Thousands of exons are found to be significantly changing including a number of exons in known estrogen responsive genes such as myc, the BCL-1 oncogene, and the Stromal cell-derived factor 1 precursor (SDF-1). Interestingly, these genes are generally involved in regulating proliferation. Thus, we are building a network of genes and exons that may be part of a breast cancer cell's response to estrogen. Finally, this network will be compared to the ChIP-chip data of ER and RNA binding proteins to understand the regulation of gene expression at the post-transcriptional level in response to estrogen.

KEY RESEARCH ACCOPLISHMENTS

- Developed the chromatin IP approach to localize RNA binding proteins to the human genome.
- Identified new gene targets of RNA binding proteins in the ENCODE regions.
- Representative splicing factors PTB and U2AF65 are recruited to many genes at the 5' end but have exon specificity within the gene.
- Localized two states of RNA Polymerase II to the ENCODE regions in HeLa cells.

- A major regulator of transcription elongation is coupling to pre-mRNA processing.
- Identified novel sites of transcription in the ENCODE regions
- Localized RNA Polymerase II to the whole-genome in breast cancer cells.
- Identifying novel sites of transcription in response to estrogen.
- Discovered novel sites of Estrogen Receptor regulation across the human genome including potential microRNAs.
- Discovering novel genes that respond to estrogen at the level of alternative splicing.

REPORTABLE OUTCOMES

Carroll, J. S., Liu, X. S., Brodsky, A. S., Li, W., Szary, A. J., Shao, W., Meyer, C. A., Hestermann, E. V., Geistlinger, T. R., Fox, E. A., Silver, P. A., Brown, M. “Chromosome-wide Mapping of Estrogen Receptor Binding Reveals Long-range Combinatorial Regulation” *submitted..*

Brodsky, A. S., Meyer, C. A., Hall, G., Swinburne, I., Keenan, B. K., Liu, X., Fox, E. A., and Silver, P. A. “Genomic Mapping of RNA Polymerase II Reveals Sites of Co-Transcriptional Regulation in Human Cells” *submitted.*

Brodsky, A.S., Johnston, A.P., Trau, M., and Silver P.A. “Analysis of RNA-protein interactions by flow cytometry” *Curr Opin Mol Ther.* **5** 235-40 (2003).

Brodsky, A. S._and Silver, P. A. “A microbead-based system for identifying and characterizing RNA-protein interactions by flow cytometry” *Mol Cell Proteomics*, **1** 922-929 (2002).

All microarray data files are being deposited in the Gene Expression Omnibus (GEO) database.

CONCLUSIONS

We have developed an approach to examine the role of RNA binding proteins in post-transcriptional regulation in human cells. We have applied this technology to understand the post-transcriptional regulation of important oncogenes such as MDM2 as well as the cell’s response to estrogen. A number of groups have found that the mRNA levels of only ~100 genes change upon stimulation of estrogen in breast cancer cells [9]. Furthermore, these are no clear patterns amongst these 100 genes to understand how the cell is responding to estrogen. Thus, we have developed a platform to examine the role of RNA binding proteins and post-transcriptional regulation in breast cancer cells. The role of RNA binding proteins in cancer remains unclear but learning which RNA binding proteins may be regulating oncogenes and estrogen responsive genes may provide clues to the role of post-transcriptional mechanisms. We have learned that RNA Polymerase accumulates at exons across genes and are identifying novel sites of transcription in breast cancer. These include regions around microRNAs and other noncoding RNAs.

Thus, during this grant funding period, we have developed genomic approaches to localize RNA binding proteins to the genome and measure post-transcriptional responses. We have begun learning about the role of RNA binding proteins and post-transcriptional regulation in breast cancer cells. These data are providing new insights into the regulation of gene expression in breast cancer cells.

REFERENCES

1. Brodsky AS, Silver PA: **A microbead-based system for identifying and characterizing RNA-protein interactions by flow cytometry.** *Mol Cell Proteomics* 2002, **1**(12):922-929.
2. Brodsky AS, Johnston AP, Trau M, Silver PA: **Analysis of RNA-protein interactions by flow cytometry.** *Curr Opin Mol Ther* 2003, **5**(3):235-240.
3. Lei EP, Silver PA: **Intron status and 3'-end formation control cotranscriptional export of mRNA.** *Genes Dev* 2002, **16**(21):2761-2766.
4. Lei EP, Krebber H, Silver PA: **Messenger RNAs are recruited for nuclear export during transcription.** *Genes Dev* 2001, **15**(14):1771-1782.
5. Ostermeier GC, Liu Z, Martins RP, Bharadwaj RR, Ellis J, Draghici S, Krawetz SA: **Nuclear matrix association of the human beta-globin locus utilizing a novel approach to quantitative real-time PCR.** *Nucleic Acids Res* 2003, **31**(12):3257-3266.
6. Bartel F, Taubert H, Harris LC: **Alternative and aberrant splicing of MDM2 mRNA in human cancer.** *Cancer Cell* 2002, **2**(1):9-15.
7. **The ENCODE (ENCyclopedia Of DNA Elements) Project.** *Science* 2004, **306**(5696):636-640.
8. Michelotti EF, Michelotti GA, Aronsohn AI, Levens D: **Heterogeneous nuclear ribonucleoprotein K is a transcription factor.** *Mol Cell Biol* 1996, **16**(5):2350-2360.
9. Lin CY, Strom A, Vega VB, Kong SL, Yeo AL, Thomsen JS, Chan WC, Doray B, Bangarusamy DK, Ramasamy A *et al*: **Discovery of estrogen receptor alpha target genes and response elements in breast tumor cells.** *Genome Biol* 2004, **5**(9):R66.

PERSONNEL RECEIVING PAY FROM THE RESEARCH EFFORT:

Pamela A. Silver, Ph.D., Principal Investigator

Alexander Brodsky, Ph.D., Research Fellow

Cheryl Wojciechowski, Ph.D., Research Fellow

Vicki Losick, Research Technician

Benjamin Keenan, Research Technician

APPENDIX

Contents: Copies of all publications resulting from this research effort:

1. Carroll, J. S., Liu, X. S., Brodsky, A. S., Li, W., Szary, A. J., Shao, W., Meyer, C. A., Hestermann, E. V., Geistlinger, T. R., Fox, E. A., Silver, P. A., Brown, M. "Chromosome-wide Mapping of Estrogen Receptor Binding Reveals Long-range Combinatorial Regulation" *submitted..*

2. Brodsky, A. S., Meyer, C. A., Hall, G., Swinburne, I., Keenan, B. K., Liu, X., Fox, E. A., and Silver, P. A. "Genomic Mapping of RNA Polymerase II Reveals Sites of Co-Transcriptional Regulation in Human Cells" *submitted.*

3. Brodsky, A.S., Johnston, A.P., Trau, M., and Silver P.A. "Analysis of RNA-protein interactions by flow cytometry" *Curr Opin Mol Ther.* 5 235-40 (2003).

4. Brodsky, A. S._and Silver, P. A. "A microbead-based system for identifying and characterizing RNA-protein interactions by flow cytometry" *Mol Cell Proteomics*, 1 922-929 (2002).

Chromosome-wide Mapping of Estrogen Receptor Binding Reveals Long-range Combinatorial Regulation Requiring Forkhead Proteins

Mapping of estrogen receptor binding to chromosomes 21 and 22 using chromatin immunoprecipitation and tiled microarrays reveals the importance of Forkhead factors in estrogen-regulated gene expression.

Jason S. Carroll[1], Xiaole S. Liu[2,4], Alexander S. Brodsky[3,5], Clifford A. Meyer[2,4], Wei Li[2,4], Anna J. Szary[1], Jerome Eeckhoute[1], Wenlin Shao[1], Eli V. Hestermann[6], Timothy R. Geistlinger[1], Edward A. Fox[3], Pamela A. Silver[3,5] and Myles Brown[1*]

[1]Department of Medical Oncology, [2]Department of Biostatistics and [3]Department of Cancer Biology, Dana-Farber Cancer Institute, Harvard Medical School. 44 Binney St, Boston, MA, 02115. [4]Harvard School of Public Health. [5]Department of Systems Biology, Harvard Medical School. [6]Department of Biology, Furman University, 3300 Ponsett Hwy, Greenville, SC, 29613

*To whom correspondence should be addressed. E-mail: myles_brown@dfci.harvard.edu

Estrogen plays an essential physiologic role in reproduction and a pathologic one in the development and progression of breast and endometrial cancers. The completion of the human genome has allowed the identification of the expressed regions of almost all protein-coding genes, however, little is known concerning the organization of their cis-regulatory elements. We have mapped the association of the estrogen receptor (ER) with the complete non-repetitive sequence of human chromosomes 21 and 22 by combining chromatin immunoprecipitation (ChIP) with tiled microarrays. ER binds selectively to a limited number of sites, the majority of which are distant (often greater than 100 kb) from the transcription start site of regulated genes. Surprisingly, the unbiased sequence interrogation of the pool of genuine chromatin binding sites suggests that direct ER binding through canonical EREs and ERE half-sites requires in addition the presence of Forkhead factor binding in close proximity to ER binding. Furthermore, knockdown of Forkhead factor expression blocks the association of ER with chromatin and estrogen-induced gene expression demonstrating the necessity for combinatorial interaction between these two signaling pathways in mediating an estrogen response.

Introduction

Estrogen is an essential regulator of female development and reproductive function and has been implicated as a causal factor in breast and endometrial cancers. Estrogen-regulated gene expression is mediated by the action of two members of the nuclear receptor family, ERα and ERβ, with ERα being dominant in both breast epithelial cells and in breast cancer. Significant progress has been made over the past decade in defining the complex interactions between chromatin and an array of factors involved in ER-mediated gene expression (Halachmi et al., 1994; Metivier et al., 2003; Shang and Brown, 2002; Shang et al., 2000). These include the cyclic association of ER, p160 coactivators (such as AIB-1), histone acetyl transferases (HAT) and chromatin modifying molecules, such as p300/CBP and p/CAF, with target promoters in an ordered temporal fashion (Metivier et al., 2003; Shang et al., 2000).

In addition, a number of strategies including most recently gene expression profiling on microarrays have identified potential ER target genes in human breast cancer cells. Of these genes the cis-element targeted directly by ER has been identified for only a small subset. Estrogen Responsive Elements (ERE) have been identified within the 1 kb 5'-proximal region of the estrogen-regulated genes TFF-1 (pS2), EBAG9 and Cathepsin D (Augereau et al., 1994; Berry et al., 1989; Ikeda et al., 2000). The proximal promoters of target genes that lack EREs, including c-Myc and IGF-I, contain AP-1 and Sp-1 sites that appear essential for transcription in *in vitro* reporter assays (Dubik and Shiu, 1992; Umayahara et al., 1994). Few, if any regulatory elements at significant distances from the mRNA start sites of target genes have been shown to be directly targeted by ER and computation approaches to identify novel ER-binding domains have focused primarily on gene proximal regions (Bajic and Seah, 2003; Bourdeau et al., 2004).

In contrast, a wealth of studies on β-globin gene regulation has contributed to our understanding of general mechanisms of transcriptional regulation and has shown that

Locus Control Regions (LCR) up to 25 kb from the gene, are capable of enhancing gene transcription (recently reviewed in (Bulger et al., 2002)). In this study we have undertaken an unbiased approach to identify all regulatory regions that may play a role in ER-mediated transcription, by combining chromatin immunoprecipitation (ChIP) analyses of *in vivo* ER-chromatin complexes with Affymetrix tiled oligonucleotide microarrays that cover the entire non-repetitive sequences of chromosomes 21 and 22, including, importantly, all the intergenic regions. Most previous ChIP-microarrays studies have focused primarily on promoter regions (Kapranov et al., 2002; Odom et al., 2004) or CpG islands, which represent promoter-rich sequences (Weinmann et al., 2002). The tiled arrays used here are composed of 25 bp probes located at 35 nucleotide resolution (Cawley et al., 2004; Kapranov et al., 2002) and permit the opportunity to interrogate previously unexplored regions of chromosomal DNA. The 780 characterized or predicted genes on chromosomes 21 and 22 represent about 2% of the total number of genes (Kapranov et al., 2002) and thus provides a representative model for the unbiased identification of paradigms of gene regulation by ER.

Here we find a discrete number of ER binding sites across chromosomes 21 and 22, almost all of which are in non-promoter proximal regions. We explored underlying biological patterns within the list of genuine chromatin interacting domains and identified common motifs highly enriched in these regions. Using this information we prove that the distal ER binding sites are discrete chromatin regions involved in transcriptional regulation and that Forkhead proteins, at these sites, are required for activity by the ER.

Results

ER occupies a limited number of binding sites on chromosome 21 and 22

Estrogen-dependent MCF-7 breast cancer cells were deprived of hormones and stimulated with estrogen or vehicle for 45 minutes, a time we have previously shown to have maximal recruitment of ER to the promoters of several known gene targets, including Cathepsin D and TFF-1 (Shang et al., 2000). Following ChIP, ER-associated DNA was amplified using non-biased conditions, labelled and hybridized to the tiled microarrays. Relative confidence prediction scores were generated by quantile normalization across each probe followed by an analysis using a two-state Hidden Markov model (Rabiner, 1989). These scores included probe intensity and width of probe cluster. Triplicate experiments eliminated stochastic false positives after which peaks that appeared at least twice in the three replicates were included. Real time PCR primers were designed against numerous peaks in the list and directed ER ChIP was conducted to identify the boundary between the true ER binding peaks (> 1.5 fold enrichment over input) and the false positives (data not shown). Following filtering, the final list contained a total of 57 estrogen-stimulated ER binding sites within 32 discrete clusters (Fig 1A, 1B and supplemental data 1).

As one example of the validity of this method, the localization of ER to the proximal promoter 400 bp region of the estrogen-regulated gene, TFF-1, was observed. A functional ERE had been previously mapped to the region 393 to 405 bp upstream from the transcription start site of TFF-1 (Berry et al., 1989). Furthermore a region 10.5 kb upstream of the TFF-1 transcription initiation site (Fig 1A) was also found to be bound by ER. Interestingly, an estrogen-inducible DNase I hypersensitive site has been previously mapped 10.5 kb upstream from the TFF-1 start site (Giamarchi et al., 1999), though the region had not been further characterized. Our data now define this region as an authentic ER-binding site.

Within the small list of 32 ER-binding clusters, we observed interaction with a number of genes previously implicated as estrogen targets, including the transcription factor XBP-1, DSCAM-1 and the nuclear receptor co-regulator NRIP-1 (Cavailles et al., 1995; Pedram et al., 2002; Wang et al., 2004). Binding sites were also observed within 200 kb from genes not previously implicated as estrogen targets, including SOD-1, a superoxide dismutase gene involved in scavenging oxygen free radicals (Beckman et al., 1993; Singh et al., 1998) and implicated in tamoxifen-resistant progression in MCF-7 xenografts (Schiff et al., 2000). None of these genes recruited ER to a proximal 5' promoter region, but possessed divergent patterns of association. The XBP-1 gene, recruited ER to three distinct and discrete regions 13.2 kb to 22.9 kb upstream of the transcription start site (Fig 1B). DSCAM-1 contained a clustering of ten intronic ER binding sites, more than 0.5 Mb from the transcription initiation site. NRIP-1 contained four ER-binding sites in a region of chromosome 21 well known for its scarcity of genes (Katsanis et al., 1998). 5' RACE was performed on NRIP-1 to determine the exact location of the transcription start site and to identify the distance between the ER binding sites and the genuine transcriptional start site. Sequencing of the 5' terminus of the NRIP-1 transcript after estrogen stimulation revealed the presence of two previously missed exons for NRIP-1, 74.96 kb and 97.39 kb from the previously annotated gene start site (data not shown).

Therefore, the ER-binding domains exist 107 to 144 kb from the genuine transcription start site of NRIP-1. The locations of all binding sites in relation to genes can be found in supplementary data 4.

The ER binding sites adjacent to TFF-1, XBP-1, SOD-1, NRIP-1, and DSCAM-1 were validated by ER ChIP and standard PCR (Fig 2A-E). Also, quantitative PCR was performed on each of these sites after ER ChIP (Fig 2F) confirming these putative *in vivo* binding sites as genuine ER binding sites. To test whether these discrete ER recruitment regions are unique to estrogen action in MCF-7 cells, we performed ER ChIP and directed real time PCR against the same sites in T47-D breast cancer cells. These data confirm that the majority of the sites identified in MCF-7 cells are also regions of estrogen dependent ER-binding in a second ER-positive breast cancer cell line (data not shown) highlighting the conservation of specific ER-chromatin association sites.

A significant number of ER binding sites reside adjacent to estrogen gene targets

Estrogen-mediated transcript changes were identified by converting RNA from vehicle or estrogen-stimulated MCF-7 cells into double stranded cDNA and hybridizing to the chromosome 21 and 22 tiled microarrays. 35 genes (4.4% of all genes) appeared to be transcribed, after which real time primers were made against all these transcripts and quantitative RT-PCR showed that 12 transcripts on chromosomes 21 and 22 were estrogen induced (Table 1). Eleven of these 12 genes had ER binding clusters within 200 kb. The only estrogen-regulated gene that did not have an adjacent ER binding cluster was ATP5J. TFF-1, XBP-1 and NRIP-1 were in the small list of 1.5% of genes upregulated following estrogen stimulation (supplemental data 1). DSCAM-1 and SOD-1 were not upregulated by estrogen stimulation at the 3 hr time point assessed, but were transcribed after 6 hr of estrogen stimulation, as determined by RT-PCR (supplemental data 3).

The delay between ER association and transcription of DSCAM-1 and SOD-1 may be a consequence of a requirement for subsequent modification of the receptor complex or the requirement for the production of other factors involved in ER action but not necessarily part of an ER complex. Regardless of the mechanism for the transcriptional delay, it now appears that early and at least some delayed estrogen-regulated genes recruit the receptor with the same kinetics. This implies that events subsequent to ER binding are responsible for timing the initiation of transcription of these delayed targets.

Distal ER binding domains function as transcriptional enhancers

The significant sequence distance between many of the ER binding sites and the putative target gene complicates their functional validation. However, we explored the possibility that these ER binding sites may recruit components indicative of transcriptional activation. RNA PolII ChIP followed by real-time PCR was performed on a subset of the putative regulatory regions adjacent to TFF-1, XBP-1, DSCAM-1, NRIP-1 and SOD-1 genes. Interestingly, RNA PolII association was seen with all of these sites in an estrogen-dependent manner (Fig 2F). Furthermore ChIP of AIB-1, an oncogenic ER coactivator (Kuang et al., 2004; Torres-Arzayus et al., 2004), confirmed that AIB-1 is also present on all of these 'regulatory' sites following estrogen exposure (Fig 2F). As negative controls, primers were designed against the intergenic region between the TFF-1 promoter and enhancer and against a region 7 kb from XBP-1 enhancer 3. Neither ER nor any of the other factors was found associated with these control regions. In addition, we examined the promoter of XBP-1. Although ER protein association was not observed at the XBP-1 promoter, RNA PolII was found enriched at this site supporting the hypothesis that XBP-1 is transcriptionally activated by ER.

To explore the possibility that the distal enhancer regions not only function as sites of protein recruitment, but physically play a role during transcription of the adjacent gene, we performed a chromosome capture assay (Dekker et al., 2002) to assess whether promoter and enhancer sequence were components of the same chromatin regions. Hormone depleted MCF-7 cells were stimulated with vehicle or estrogen and the fixed chromatin was digested with a specific restriction enzyme (BtgI), followed by ER ChIP and ligation. After ligation, the ligated chromatin mix was washed and the cross-linking was reversed. One primer in the TFF-1 promoter and one primer in the TFF-1 enhancer were used to PCR potentially ligated fragments of DNA (Horike et al., 2005). As seen in Fig 3A, TFF-1 promoter and enhancer DNA was ligated together only in the presence of estrogen, confirming that estrogen-mediated transcription of TFF-1 involves direct physical interaction between the enhancer and promoter. No interaction was seen in the no digestion control or no ligation control. We performed the same experiment using the BsmI restriction enzyme that cuts the genuine NRIP-1 promoter (as determined by 5' RACE) and enhancer 3 region. Remarkably, after ligation, we were able to PCR a 1 kb fragment that corresponded to the annealed promoter-enhancer regions using one promoter specific and one enhancer specific primer (Fig 3B). This estrogen-dependent interaction of the distal (144 kb) ER-binding site with the promoter of the NRIP-1 gene confirms the authenticity of these distal sites as transcriptional regulatory domains.

The finding that RNA PolII is recruited to the majority of ER binding sites even those removed from known transcription sites led us to investigate the possibility that these binding sites can function as genuine enhancers. To this end, we cloned 23 ER sites (40% of all ER binding sites) into a pGL-3 luciferase vector containing an SV40 promoter and transfected these vectors into hormone depleted MCF-7 cells which where subsequently treated with estrogen or vehicle control. PGL3 empty vector was used as a negative control and transfections were normalized with pRL-Null. Almost 75% of the ER binding domains contained estrogen-induced enhancer characteristics in an in vitro transcription model (Fig 3C), supporting the hypothesis that the distal binding sites play transcriptional regulatory roles.

ER binding sites are conserved across species

To identify if the ER binding sites are conserved between human and mouse genomes, we assessed the identity in sequence in a window of 6 kb from the center of all 57 ER binding sites. This conservation was mapped within a 500 bp window at a single nucleotide resolution and confirms a strong conservation at the center of the ER binding site and the 500 bp on either side of the middle of the peak (Fig 4A). However, conservation decreased to background levels at a distance of 1 kb or more from the center of the ER binding sites. This supports the hypothesis that the discrete ER binding sites we see in MCF-7 cells are conserved between species and likely play a more general role in ER action in other cellular systems.

A screen for common sequences enriched in genuine ER-binding regions suggests the importance of Forkhead factors in estrogen action

An unbiased search for common sequence motifs (Liu et al., 2002) within the 57 individual ER-binding sites on chromosomes 21 and 22 revealed the significant recurrence of two motifs. A consensus 15 base sequence was present in 49% of all the ER binding sites on chromosomes 21 and 22 (Fig 4B) and is identical to the canonical ERE (Klinge, 2001). The likelihood of an ERE occurring in one of the ER binding sites was very significantly increased when compared to all of chromosomes 21 and 22 (p = 1.33E-15). In addition, in the ER binding sites lacking a canonical ERE, a majority were found to contain one or more ERE half-sites. The occurrence of ERE half-sites was also non-random (p = 2.16E-14). In order to confirm that our failure to find ER binding at other EREs (5,500 predicted EREs on chromosomes 21 and 22, as listed in Fig 1A and 1B) was not due to the insensitivity of the ChIP-microarray technique, we performed ChIP for ER followed by PCR for several randomly selected, predicted but non-functional perfect EREs on chromosomes 21 and 22. No ER association was found at any of these sites (data not shown).

We next determined whether DNA sequences other than the classical ERE were found at the ER binding sites by analyzing the bound sequences for conserved motifs after removing the EREs. This analysis revealed the presence of a Forkhead factor binding site in 56% of the 57 ER binding regions (Fig 4B), a finding that would only occur by chance with a probability of p = 1.23E-8. Forkhead binding motifs were found in 64% of the ER-binding regions that contain a canonical ERE. Using the consensus Forkhead motif recurring within these regions (Fig 4B), we determined the probability of this motif residing within predicted ERE regions that are not bound by ER *in vivo* (18.45%). This significant enrichment of a Forkhead motif within ER binding regions (p = 3.78E-7) suggests the presence of adjacent Forkhead motifs may play a role in determining ER binding. The proportion of the 57 ER binding sites containing EREs or Forkhead motifs is presented in Fig 4C. The finding that the largest category of sites contains both an ERE and a Forkhead motif (47.4%) strongly suggests a functional interaction.

Forkhead proteins play a combinatorial and essential role in ER binding and ER-mediated gene transcription

A combinatorial interaction between Forkhead and ER pathways has been previously suggested for a small number of specific genes. HNF-3α (FoxA1) Forkhead binding domains within the promoter of the estrogen-regulated genes TFF-1 (Beck et al., 1999) and Vitellogenin B1 (Robyr et al., 2000) have been shown to be important for gene transcription. Forkhead proteins have been shown to interact with the ER protein in yeast two hybrid experiments (Schuur et al., 2001). The function of Forkhead proteins can be regulated by their nuclear-cytoplasmic distribution depending on their phosphorylation (Brunet et al., 1999; Kops et al., 1999). We therefore assessed the localization of the Forkhead factor, FoxA1, after vehicle or estrogen stimulation of MCF-7 cells. FoxA1 protein predominantly resides in the nucleus in both vehicle and estrogen treated cells (data not shown).

We next determined whether FoxA1 was recruited along with ER to the ER-binding domains. ChIP of FoxA1, followed by real-time PCR of all 57 ER binding regions on chromosomes 21 and 22 revealed a high degree of concordance between regions that recruit ER and FoxA1. Approximately 48% of all of the ER binding domains showed FoxA1 interaction, although the pattern of recruitment differed from site to site. A majority of the regions containing FoxA1, did so in the absence of estrogen, but FoxA1 binding was decreased following estrogen stimulation. This was the case for NRIP-1 enhancer 1, DSCAM-1 enhancer 1 and TFF-1 promoter (Fig 5A). FoxA1 association with XBP-1 enhancer 2 was clearly observed, but was not diminished after estrogen addition (Fig 5A). All of these ER binding sites contained a Forkhead motif and an ERE or ERE half site (Fig 5B). FoxA1 was not seen to bind to XBP-1 enhancer 3, which lacks a Forkhead motif (Fig 5). However, several regions containing Forkhead motifs did not recruit FoxA1 and several ER binding domains that lacked Forkhead motifs did bind FoxA1. This complex interplay between FoxA1, ER and binding sites within chromatin likely involves adjacent regions to the ER binding sites and may involve other proteins. Despite this, it is clear that a significant proportion of ER binding sites, especially those adjacent to actively transcribed genes contain FoxA1 prior estrogen stimulation and ER recruitment to the same regions.

To determine the importance of FoxA1 in mediating ER association with chromatin, we developed siRNA to the 3'UTR of FoxA1 mRNA. Specific targeted knockdown of FoxA1 protein was achieved (Fig 6A), without changes in control protein or ER protein levels (data not shown). A luciferase siRNA (siLuc) was used as a negative control. MCF-7 cells were deprived of hormones for 24 hr and siLuc, or siRNA to FoxA1 was transfected for 6 hr, after which hormone depleted media was added for a further 48 hr and cells were stimulated with estrogen or vehicle. ER ChIP and real time PCR of a number of previously validated binding sites was performed. The decrease in FoxA1 completely impeded the ability of ER to bind to TFF-1 promoter, XBP-1 enhancer 1 and NRIP-1 enhancer 2 (Fig 6B), as well as DSCAM-1 enhancer 1 (data not shown). No changes were observed on the XBP-1 promoter, which functioned as a negative control (Fig 6B).

Since the targeted knockdown of FoxA1 inhibited the ability of ER to associate with in vivo ER binding sites, we assessed the effect of Forkhead down-regulation on estrogen-mediated transcription. After siLuc or siFoxA1 transfection, cells were stimulated with estrogen or vehicle for 6 hr and mRNA changes in all 12 estrogen target genes on chromosomes 21 and 22 were assessed. The estrogen-induced increases in all 12 estrogen targets were abolished when FoxA1 was down-regulated (Fig 6C), but no changes were observed in GAPDH control mRNA levels. The essential role for the FoxA1 Forkhead protein during transcription of all estrogen target genes on chromosomes 21 and 22, confirms a general requirement of FoxA1 for ER transcription.

Discussion

A complete picture of ER-mediated gene activation has begun to emerge in recent years, with a coordinated and timely cycling of receptor, nuclear coactivators, chromatin remodelling proteins and the transcription machinery on and off target promoters (Metivier et al., 2003; Shang et al., 2000). However, these studies oversimplify the problem by focusing on the promoter proximal region of one or two target genes and largely ignore the remaining chromosomal sequence. Here we have interrogated the association of ER across entire chromosomes, including intergenic regions that contain potential cis-regulatory domains. These ChIP-microarray experiments demonstrate the ability to identify genuine *in vivo* ER protein binding sites in previously unexplored regions of the genome. Interestingly, while a few of the ER binding sites were found directly adjacent to ER target genes, most were found at significant distances including several >100 kb removed from transcription start sites. Of the 57 ER binding sites (within 32 potential transcriptional regulatory clusters), only a very small number of proximal promoters recruited ER, despite the fact that other genes were estrogen induced. The presence of multiple components of the transcriptional machinery at the distal sites and the ability of chromosome conformation capture assays to demonstrate that these distant sites are physically associated with promoter-proximal regions suggests that they play an important role in estrogen-mediated regulation.

A significant volume of work has focused on identifying essential domains within the proximal promoters of known estrogen regulated genes using *in vitro* methods (Dubik and Shiu, 1992; Petz et al., 2002; Porter et al., 1996; Teng et al., 1992; Umayahara et al., 1994; Vyhlidal et al., 2000; Weisz and Rosales, 1990). The conclusions drawn from this large volume of data implicate a number of motifs, including Sp1, AP-1 and GC rich regions as important cis-regulatory domains in ER-mediated transcription. However, our data demonstrate ER regulatory sites at distances several orders of magnitude greater than was focused on in the past, suggesting that they may function in ways analogous to the β-globin LCR which has its major effect subsequent to PIC formation (Sawado et al., 2003).

Non-biased motif scanning of the genuine *in vivo* ER binding sites identified a canonical estrogen responsive element (ERE) in the majority of ER binding sites that represented only 1.5% of EREs predicted by bioinformatics alone. Previous approaches for motif identification involved computational based methods for identifying response elements, after which gene proximal sites are included as potential binding domains (Bajic and Seah, 2003; Bourdeau et al., 2004). The current data suggest that while ER binding involves interaction with consensus ERE motifs, the presence of such motifs is insufficient to dictate receptor-chromatin association. Furthermore, the exclusion of response elements further than several kilobases from transcription start sites eliminates distal regulatory regions that may be the primary receptor-chromatin interaction sites.

Since the presence of an ERE alone is insufficient to define an authentic ER regulatory site, we searched for other conserved sequences and found that Forkhead factor binding sites are present near authentic EREs significantly more frequently than those that do not bind ER. We showed that a Forkhead factor (FoxA1) binding was essential for ER-chromatin interactions and subsequent expression of estrogen gene targets. FoxA1 protein can bind condensed chromatin via its winged-helix DNA binding domains that mimic histone linker proteins (Cirillo et al., 2002; Cirillo et al., 1998). Unlike histone proteins however, FoxA1 does not contain the amino acid composition to condense chromatin and it therefore thought to promoter euchromatic conditions. As such, it is possible that the presence of FoxA1 identifies specific regions within chromatin to facilitate the association of the ER transcription complex. Our data suggest that FoxA1 is present on the chromatin at a number of regions, after which ER can associate with these specific sites. Down-regulation of FoxA1 inhibits the ability of ER to associate with its binding sites, confirming the requirement for Forkhead directed association of ER with chromatin, despite the fact that these sites contain sufficient information, in the form of an ERE, for ER docking. A recent investigation has shown that FoxA1 can directly modulate chromatin in the MMTV promoter and can positively enhance transcription by the Glucocorticoid Receptor (Holmqvist et al., 2005), supporting a general model for FoxA1 involvement in nuclear receptor transcription.

We have taken an unbiased approach to identify regions of chromatin, both promoter proximal and intergenic sequences, that are involved in ER-mediated transcriptional activity. We find a limited number of bone fide ER binding sites on chromosomes 21 and 22, with a significant enrichment of canonical ERE palindromes and half sites within the binding sites. Moreover, the presence of Forkhead binding motifs and the subsequent identification of a functional role for Forkhead proteins, exemplifies the power of this methodological approach to identify important regulatory domains within the vast regions of unexplored sequence in the human genome.

Materials and Methods

Chromatin Immunoprecipitation (ChIP)-microarray preparation

ChIP was performed as previously described (Shang et al., 2000), with the following modifications. 2 μg of antibody was prebound for a minimum of 4 hr to protein A and protein G Dynal magnetic beads (Dynal Biotech, Norway) and washed three times with ice-cold PBS plus 5% BSA, and then added to the diluted chromatin and immunoprecipitated overnight. The magnetic bead-chromatin complexes were collected and washed 6 times in RIPA buffer (50 mM HEPES pH 7.6, 1 mM EDTA, 0.7% Na deoxycholate, 1% NP-40, 0.5 M LiCl). Elution of the DNA from the beads was as previously described (Shang et al., 2000). Antibodies used were: ERα (Ab-10) from Neomarkers (Lab Vision, UK), ERα (HC-20), RNA PolII (H-224), AIB-1/RAC3 (C-20), TFIID (SI1), HNF-3α/FoxA1 (H-120), mouse IgG (sc-2025) and rabbit IgG (sc-2027) from Santa Cruz (Santa Cruz Biotechnologies, CA). Ligation-Mediated PCR was performed as previously described (Kapranov et al., 2002).

Data Analysis

1,054,325 probe pairs were mapped to chromosome 21 and 22 according to the NCBIv33 GTRANS Libraries provided by Affymetrix. (PM-MM) value was recorded for each probe pair, and a probe pair was removed if either PM or MM was flagged as outlier by the Affymetrix GCOS software. The five samples (three ER+ ChIP and three genomic inputs) were normalized by quantile normalization (Bolstad et al., 2003) based on a combined 76 ChIP experiments obtained from public domain and Dana-Farber Cancer Institute. The behavior of every probe pair *i*, assumed to be $N(\mu_i, \sigma_i^2)$, was estimated from the 76 normalized experiments. A two-state (ChIP-enriched state and non-enriched state) Hidden Markov Model with the following parameters was applied to each sample to estimate the probability of ChIP-enrichment at each probe pair location:

Transition probabilities: $\frac{300}{1{,}054{,}325}$ for transition to a different state

$1 - \frac{300}{1{,}054{,}325}$ for staying in the same state

Emission probabilities: $N(\mu_i, \sigma_i^2)$ for non-enriched hidden state

$N(\mu_i + 2\sigma_i, (1.5\sigma_i)^2)$ for enriched hidden state

To combine the results from the six samples, an enrichment score was calculated as the average enrichment probability in the three ER+ ChIP samples subtracted by the average enrichment probability in the two genomic input samples. Since the tiling array has one 25-mer probe in every 35 bp of non-repeat regions, the coverage of every probe was extended by 10-bp on both ends. An enriched regions is defined as run of probes with enrichment score > 50% and covering at least 125 bp. Each enriched region can tolerate up to two neighboring probes with enrichment score between [10%, 50%]. If two neighboring probes are more than 210 bp apart, the enriched region is broken into two separate blocks. A summary enrichment score was obtain for each enriched region, which is the enrichment score summation for all the probes in the region divided by the square root of the number of probes in the region. This summary enrichment score represents the relative confidence of a predicted enriched region.

The genomic DNA of every ChIP-enriched region was retrieved from UCSC genome browser, and ranked by the summary enrichment score. MDscan algorithm (Liu et al., 2002) was applied to the sequences to find enriched sequence pattern that is the putative estrogen receptor binding motif. To find a motif of width *w*, MDscan first enumerates each *w*-mer in the highest ranking sequences, and collects other *w*-mers similar to it in these sequences to construct a candidate motif as a probability matrix. A semi-Bayes scoring function was used to remove low scoring candidate motifs, and refine the rest by checking all *w*-mers in all the ChIP-enriched sequences. A high scoring motif (with similar consensus) consistently reported multiple times at different motif widths indicates a strong prediction.

Species Conservation

We expanded all 57 of the ER binding sites equally in each direction to have a length 6 kb. The human-mouse conservation score of each nucleotide in the expanded binding region is defined as the average sequence identity (#Matched Nucleotides-#Indels/500) of a 500-mer window centered at the nucleotide. The human (hg15) /mouse(mm3) BLASTZ (Schwartz et al., 2003) genome alignments were downloaded from http://genome.ucsc.edu.

Real time PCR

Primers were selected using Primer Express (Applied Biosystems). 5 μl of precipitated and purified DNA was subjected to PCR using the Applied Biosystems SYBR Green Mastermix. Relative DNA quantities were measured using the PicoGreen system (Molecular Probes, OR). All primer sequences and locations are listed in supplemental data 2.

Double stranded cDNA synthesis

Total RNA was converted to double stranded cDNA according to the In Vitrogen Superscript Double stranded cDNA synthesis manufacturer's instructions. The RNA was primed with 250 ng oligo(dT) (In Vitrogen) and 25 ng random hexamers (Gibco). cDNA was fragmented and labelled as described above.

5'RACE

5' RACE was performed according to the manufacturer's instructions (In Vitrogen). The primers sequences used were: NRIP-1 RT primer (5'-TGCCTGATGCATTAGTAATCC-3'), NRIP-1 nested primer 1 (5'-GAGCCAAGCTCTTCTCCATGAGTCATGTTC-3') and NRIP-1 nested primer 2 (5'-ACCTTCCATCGCAATCAGAGAGAGACGTACTG-3'). The PCR product was cloned and sequenced by standard methods.

Chromosome capture assay

Fixed chromatin was digested overnight with specific restriction enzymes after which ER ChIP was set up as described above. After overnight ChIP, the beads were precipitated and resuspended in ligation buffer (NEB, MA, USA) and overnight ligation was performed. The beads were collected, washed and the formaldehyde cross-linking was reversed as described above. Primers used to amplify annealed fragments were as described in supplemental data 2.

Luciferase transcriptional activity

ER binding sites were amplified by PCR and cloned into the pGL3-promoter vector (Promega). Hormone depleted MCF-7 cells were transfected with each of the ER binding domain vectors with Lipofectamine 2000 (Invitrogen) and total protein lysate was harvested after estrogen or ethanol addition for 24 hr. Transfections were normalized by the co-transfection of the pRL-Null Renilla luciferase vector and Renilla and Firefly luciferase activity was assessed using the dual luciferase kit (Promega).

Western blotting

SDS-PAGE was performed as previously described (Carroll et al., 2000). Antibodies used were HNF-3α (ab5089), from AbCam (Cambridge, UK) and Calnexin (H-70) from Santa Cruz (CA, USA).

Short interfering (si) RNA

A 21 bp siRNA was designed against the FoxA1 transcript and synthesized by Dharmacon (Lafayette, CO). siRNA was transfected using Lipofectamine 2000 (InVitrogen). The siRNA sequences used were: siFoxA1 sense 5'-GAGAGAAAAAUCAACAGC-3' and antisense 5'-GCUGUUGAUUUUUCUCUC -3'; siLuc sense 5'- CACUUACGCUGAGUACUUCGA –3' and antisense 5'-UCGAAGUACUCAGCGUAAGUG –3'.

Acknowledgments

The authors would like to thank Danielle Lynch for reviewing the manuscript. We thank Molin Wang, David Harrington, Cheng Li, and Jun S. Liu for their insights and advices on the tiling array data analysis. This work was supported by grants from DOD grant DAMD17-01-1-0222 (M.B), DOD grant W81XWH-04-1-0512 (J.S.C), DOD grant DAMD17-02-0364 (P.A.S) and NHGRI grant K22 HG02488-01A1 (A.S.B).

Figure legends

Fig. 1

Map of ER binding sites on chromosomes 21 and 22 after estrogen stimulation. The visual representation of ER-binding sites on chromosomes 21 **(A)** and 22 **(B)** are shown. Genes locations are shown in blue bars. Gene locations are based on the 2003 genome freeze in the UCSC browser using Genbank RefSeq positions. Predicted EREs are shown as black bars and ER-binding sites are shown as red bars. **(A)** An expanded view of the TFF-1 gene region is shown as signal difference between ER ChIP and Input DNA for both the estrogen and vehicle treated cells. The TFF-1 gene is shown in its genuine 3'-5' orientation. The gene adjacent to TFF-1 is not an estrogen target. **(B)** Expanded view of the XBP-1 gene region on chromosome 22. The XBP-1 gene is shown in its genuine 3'-5' orientation.

Fig. 2

Validation of the *in vivo* binding of the transcription complex to regulatory regions. ChIP of ER and standard PCR of sites adjacent to TFF-1 **(A)**, XBP-1 **(B)**, DSCAM-1 **(C)**, NRIP-1 **(D)** and SOD-1 **(E)**. TFF-1 non-specific and XBP-1 promoter primers were included as negative controls. The lanes are vehicle (V), estrogen (E) and Input (I). **(F)** ChIP of ER, RNA PolII, AIB-1 or IgG control and real-time PCR of the enhancer regions. The data are estrogen-mediated fold enrichment compared to vehicle (ethanol) control and are the average of three separate replicates. The color intensity reflects the fold change as described in the legend. TFF-1 non-specific and XBP-1 non-specific primers were included as negative controls.

Table 1

List of ER binding site clusters and relative locations to putative gene targets. The 32 transcriptional clusters are shown, with the start and stop locations of the ER binding sites.

Fig. 3

Interaction of promoter-enhancer domains and transcriptional activity of enhancer regions. **(A)** Chromosome capture assay was performed after digesting fixed chromatin from vehicle or estrogen treated cells with the BtgI restriction enzyme. Primers flanking the TFF-1 promoter and enhancer were used to amplify DNA after ligation. Undigested controls and no ligase controls were included. **(B)** Chromatin was digested with BsmI and one primer flanking the NRIP-1 promoter and one in enhancer 3 region were used to amplify a specific product after ligation. **(C)** ER binding sites were cloned into the pGL-3 promoter vector and transfected into hormone depleted MCF-7 cells, after which vehicle or estrogen was added. Empty pGL3-promoter vector was used as a negative control. Co-transfection of pRL-Null Renilla vector was included as a normalizing control. The open bars represent vehicle treated and black bars represent estrogen treatment.

Fig. 4

Conservation of ER binding sites and presence of enriched motifs. **(A)** Sequence homology of ER binding sites and surrounding sequence between human and mouse genomes. The center of ER peaks is designated co-ordinate 0. **(B)** An unbiased motif screen of all the ER binding sites on chromosomes 21 and 22 revealed the presence of two enriched motifs, an ERE and a Forkhead binding motif, both of which are visually represented in WebLogo (http://weblogo.berkeley.edu). **(C)** The occurrence of ERE or ERE half sites and Forkhead sites within the 57 ER binding sites on chromosomes 21 and 22.

Fig. 5

Recruitment of Forkhead protein to ER-binding domains. **(A)** ChIP of FoxA1 followed by real-time PCR of NRIP-1 enhancer 1, DSCAM-1 enhancer 1, TFF-1 promoter and XBP-1 enhancer 1. XBP-1 enhancer 2 is included as a control which does not recruit

FoxA1. Data is shown as fold change versus input. **(B)** Schematic diagram showing the relative location of ERE motifs (inverted green arrows), ERE half-sites (blue arrows) and Forkhead motifs (red arrows). Chromosome nucleotide locations are given.

Fig. 6

Specific targeted knockdown of FoxA1 protein and the effects on estrogen-mediated transcription. **(A)** siRNA to FoxA1 was transfected into hormone depleted MCF-7 cells and changes in protein levels were determined after vehicle or estrogen treatment. SiLuc was used as a transfection control and Calnexin was used as a loading control. **(B)** ER ChIP was performed after vehicle or estrogen treatment of siLuc or siFoxA1 transfected cells and real time PCR was conducted on TFF-1 promoter, XBP-1 enhancer 1, NRIP-1 enhancer 2 and XBP-1 enhancer 2 as a negative control. The data are fold enrichment over vehicle treated. **(C)** Changes in mRNA levels of all estrogen-regulated genes on chromosomes 21 and 22. The data are estrogen-mediated fold enrichment compared to vehicle (ethanol) control and are the average of three separate replicates. The color intensity reflects the fold change as described in the legend.

References

Augereau, P., Miralles, F., Cavailles, V., Gaudelet, C., Parker, M., and Rochefort, H. (1994). Characterization of the proximal estrogen-responsive element of human cathepsin D gene. Mol Endocrinol *8*, 693-703.

Bajic, V. B., and Seah, S. H. (2003). Dragon Gene Start Finder identifies approximate locations of the 5' ends of genes. Nucleic Acids Res *31*, 3560-3563.

Beck, S., Sommer, P., dos Santos Silva, E., Blin, N., and Gott, P. (1999). Hepatocyte nuclear factor 3 (winged helix domain) activates trefoil factor gene TFF1 through a binding motif adjacent to the TATAA box. DNA Cell Biol *18*, 157-164.

Beckman, J. S., Carson, M., Smith, C. D., and Koppenol, W. H. (1993). ALS, SOD and peroxynitrite. Nature *364*, 584.

Berry, M., Nunez, A. M., and Chambon, P. (1989). Estrogen-responsive element of the human pS2 gene is an imperfectly palindromic sequence. Proc Natl Acad Sci U S A *86*, 1218-1222.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185-193.

Bourdeau, V., Deschenes, J., Metivier, R., Nagai, Y., Nguyen, D., Bretschneider, N., Gannon, F., White, J. H., and Mader, S. (2004). Genome-wide identification of high-affinity estrogen response elements in human and mouse. Mol Endocrinol *18*, 1411-1427.

Brunet, A., Bonni, A., Zigmond, M. J., Lin, M. Z., Juo, P., Hu, L. S., Anderson, M. J., Arden, K. C., Blenis, J., and Greenberg, M. E. (1999). Akt promotes cell survival by phosphorylating and inhibiting a Forkhead transcription factor. Cell *96*, 857-868.

Bulger, M., Sawado, T., Schubeler, D., and Groudine, M. (2002). ChIPs of the beta-globin locus: unraveling gene regulation within an active domain. Curr Opin Genet Dev *12*, 170-177.

Carroll, J. S., Prall, O. W., Musgrove, E. A., and Sutherland, R. L. (2000). A pure estrogen antagonist inhibits cyclin E-Cdk2 activity in MCF-7 breast cancer cells and induces accumulation of p130-E2F4 complexes characteristic of quiescence. J Biol Chem *275*, 38221-38229.

Cavailles, V., Dauvois, S., L'Horset, F., Lopez, G., Hoare, S., Kushner, P. J., and Parker, M. G. (1995). Nuclear factor RIP140 modulates transcriptional activation by the estrogen receptor. Embo J *14*, 3741-3751.

Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., *et al.* (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell *116*, 499-509.

Cirillo, L. A., Lin, F. R., Cuesta, I., Friedman, D., Jarnik, M., and Zaret, K. S. (2002). Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. Mol Cell *9*, 279-289.

Cirillo, L. A., McPherson, C. E., Bossard, P., Stevens, K., Cherian, S., Shim, E. Y., Clark, K. L., Burley, S. K., and Zaret, K. S. (1998). Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome. Embo J *17*, 244-254.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306-1311.

Dubik, D., and Shiu, R. P. (1992). Mechanism of estrogen activation of c-myc oncogene expression. Oncogene *7*, 1587-1594.

Giamarchi, C., Solanas, M., Chailleux, C., Augereau, P., Vignon, F., Rochefort, H., and Richard-Foy, H. (1999). Chromatin structure of the regulatory regions of pS2 and cathepsin D genes in hormone-dependent and -independent breast cancer cell lines. Oncogene *18*, 533-541.

Halachmi, S., Marden, E., Martin, G., MacKay, H., Abbondanza, C., and Brown, M. (1994). Estrogen receptor-associated proteins: possible mediators of hormone-induced transcription. Science *264*, 1455-1458.

Holmqvist, P.-H., Belikov, S., Zaret, K. S., and Wrange, Ö. (2005). FoxA1 bnding to the MMTV LTR modulates chromatin structure and transcription. Exp Cell Res *In Press*.

Horike, S., Cai, S., Miyano, M., Cheng, J. F., and Kohwi-Shigematsu, T. (2005). Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. Nat Genet *37*, 31-40.

Ikeda, K., Sato, M., Tsutsumi, O., Tsuchiya, F., Tsuneizumi, M., Emi, M., Imoto, I., Inazawa, J., Muramatsu, M., and Inoue, S. (2000). Promoter analysis and chromosomal mapping of human EBAG9 gene. Biochem Biophys Res Commun *273*, 654-660.

Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P., and Gingeras, T. R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. Science *296*, 916-919.

Katsanis, N., Ives, J. H., Groet, J., Nizetic, D., and Fisher, E. M. (1998). Localisation of receptor interacting protein 140 (RIP140) within 100 kb of D21S13 on 21q11, a gene-poor region of the human genome. Hum Genet *102*, 221-223.

Klinge, C. M. (2001). Estrogen receptor interaction with estrogen response elements. Nucleic Acids Res *29*, 2905-2919.

Kops, G. J., de Ruiter, N. D., De Vries-Smits, A. M., Powell, D. R., Bos, J. L., and Burgering, B. M. (1999). Direct control of the Forkhead transcription factor AFX by protein kinase B. Nature *398*, 630-634.

Kuang, S. Q., Liao, L., Zhang, H., Lee, A. V., O'Malley, B. W., and Xu, J. (2004). AIB1/SRC-3 deficiency affects insulin-like growth factor I signaling pathway and suppresses v-Ha-ras-induced breast cancer initiation and progression in mice. Cancer Res *64*, 1875-1885.

Liu, X. S., Brutlag, D. L., and Liu, J. S. (2002). An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat Biotechnol *20*, 835-839.

Metivier, R., Penot, G., Hubner, M. R., Reid, G., Brand, H., Kos, M., and Gannon, F. (2003). Estrogen receptor-alpha directs ordered, cyclical, and combinatorial recruitment of cofactors on a natural target promoter. Cell *115*, 751-763.

Odom, D. T., Zizlsperger, N., Gordon, D. B., Bell, G. W., Rinaldi, N. J., Murray, H. L., Volkert, T. L., Schreiber, J., Rolfe, P. A., Gifford, D. K., *et al.* (2004). Control of pancreas and liver gene expression by HNF transcription factors. Science *303*, 1378-1381.

Pedram, A., Razandi, M., Aitkenhead, M., Hughes, C. C., and Levin, E. R. (2002). Integration of the non-genomic and genomic actions of estrogen. Membrane-initiated signaling by steroid to transcription and cell biology. J Biol Chem *277*, 50768-50775.

Petz, L. N., Ziegler, Y. S., Loven, M. A., and Nardulli, A. M. (2002). Estrogen receptor alpha and activating protein-1 mediate estrogen responsiveness of the progesterone receptor gene in MCF-7 breast cancer cells. Endocrinology *143*, 4583-4591.

Porter, W., Wang, F., Wang, W., Duan, R., and Safe, S. (1996). Role of estrogen receptor/Sp1 complexes in estrogen-induced heat shock protein 27 gene expression. Mol Endocrinol *10*, 1371-1378.

Rabiner, L. R. (1989). A tutorial on Hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE *77*, 257-286.

Robyr, D., Gegonne, A., Wolffe, A. P., and Wahli, W. (2000). Determinants of vitellogenin B1 promoter architecture. HNF3 and estrogen responsive transcription within chromatin. J Biol Chem *275*, 28291-28300.

Sawado, T., Halow, J., Bender, M. A., and Groudine, M. (2003). The beta -globin locus control region (LCR) functions primarily by enhancing the transition from transcription initiation to elongation. Genes Dev *17*, 1009-1018.

Schiff, R., Reddy, P., Ahotupa, M., Coronado-Heinsohn, E., Grim, M., Hilsenbeck, S. G., Lawrence, R., Deneke, S., Herrera, R., Chamness, G. C., *et al.* (2000). Oxidative stress and AP-1 activity in tamoxifen-resistant breast tumors in vivo. J Natl Cancer Inst *92*, 1926-1934.

Schuur, E. R., Loktev, A. V., Sharma, M., Sun, Z., Roth, R. A., and Weigel, R. J. (2001). Ligand-dependent interaction of estrogen receptor-alpha with members of the forkhead transcription factor family. J Biol Chem *276*, 33554-33560.

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D., and Miller, W. (2003). Human-mouse alignments with BLASTZ. Genome Res *13*, 103-107.

Shang, Y., and Brown, M. (2002). Molecular determinants for the tissue specificity of SERMs. Science *295*, 2465-2468.

Shang, Y., Hu, X., DiRenzo, J., Lazar, M. A., and Brown, M. (2000). Cofactor dynamics and sufficiency in estrogen receptor-regulated transcription. Cell *103*, 843-852.

Singh, R. J., Karoui, H., Gunther, M. R., Beckman, J. S., Mason, R. P., and Kalyanaraman, B. (1998). Reexamination of the mechanism of hydroxyl radical adducts formed from the reaction between familial amyotrophic lateral sclerosis-associated Cu,Zn superoxide dismutase mutants and H2O2. Proc Natl Acad Sci U S A *95*, 6675-6680.

Teng, C. T., Liu, Y., Yang, N., Walmer, D., and Panella, T. (1992). Differential molecular mechanism of the estrogen action that regulates lactoferrin gene in human and mouse. Mol Endocrinol *6*, 1969-1981.

Torres-Arzayus, M. I., De Mora, J. F., Yuan, J., Vazquez, F., Bronson, R., Rue, M., Sellers, W. R., and Brown, M. (2004). High tumor incidence and activation of the PI3K/AKT pathway in transgenic mice define AIB1 as an oncogene. Cancer Cell *6*, 263-274.

Umayahara, Y., Kawamori, R., Watada, H., Imano, E., Iwama, N., Morishima, T., Yamasaki, Y., Kajimoto, Y., and Kamada, T. (1994). Estrogen regulation of the insulin-like growth factor I gene transcription involves an AP-1 enhancer. J Biol Chem *269*, 16433-16442.

Vyhlidal, C., Samudio, I., Kladde, M. P., and Safe, S. (2000). Transcriptional activation of transforming growth factor alpha by estradiol: requirement for both a GC-rich site and an estrogen response element half-site. J Mol Endocrinol *24*, 329-338.

Wang, D. Y., Fulthorpe, R., Liss, S. N., and Edwards, E. A. (2004). Identification of estrogen-responsive genes by complementary deoxyribonucleic acid microarray and

characterization of a novel early estrogen-induced gene: EEIG1. Mol Endocrinol *18*, 402-411.

Weinmann, A. S., Yan, P. S., Oberley, M. J., Huang, T. H., and Farnham, P. J. (2002). Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. Genes Dev *16*, 235-244.

Weisz, A., and Rosales, R. (1990). Identification of an estrogen response element upstream of the human c-fos gene that binds the estrogen receptor and the AP-1 transcription factor. Nucleic Acids Res *18*, 5097-5106.

Figure 1

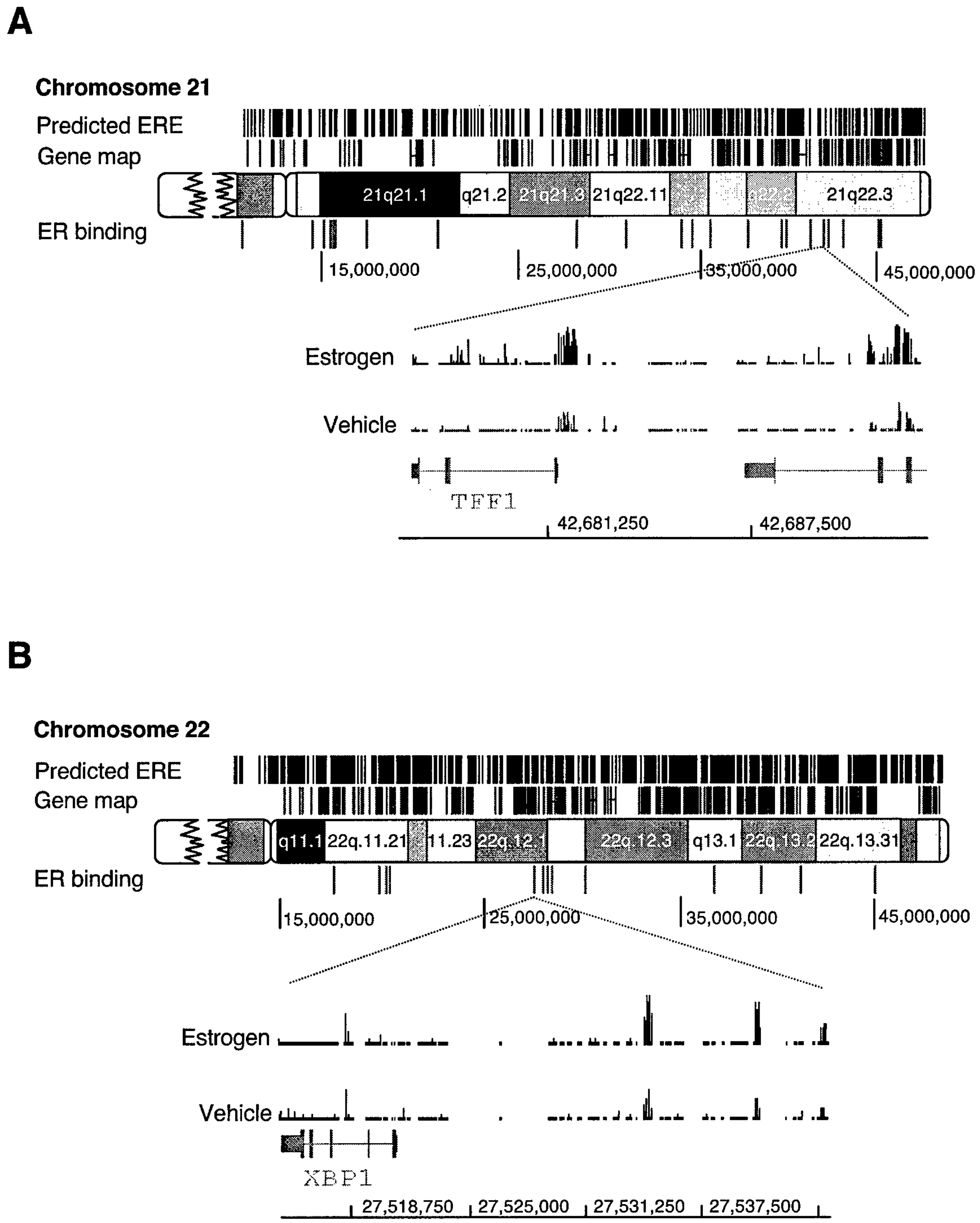

A
Chromosome 21
Predicted ERE
Gene map
21q21.1
q21.2
21q21.3
21q22.11
q22.2
21q22.3
ER binding
15,000,000
25,000,000
35,000,000
45,000,000
Estrogen
Vehicle
TFF1
42,681,250
42,687,500
B
Chromosome 22
Predicted ERE
Gene map
q11.1
22q.11.21
11.23
22q.12.1
22q.12.3
q13.1
22q.13.2
22q.13.31
ER binding
15,000,000
25,000,000
35,000,000
45,000,000
Estrogen
Vehicle
XBP1
27,518,750
27,525,000
27,531,250
27,537,500

Figure 2

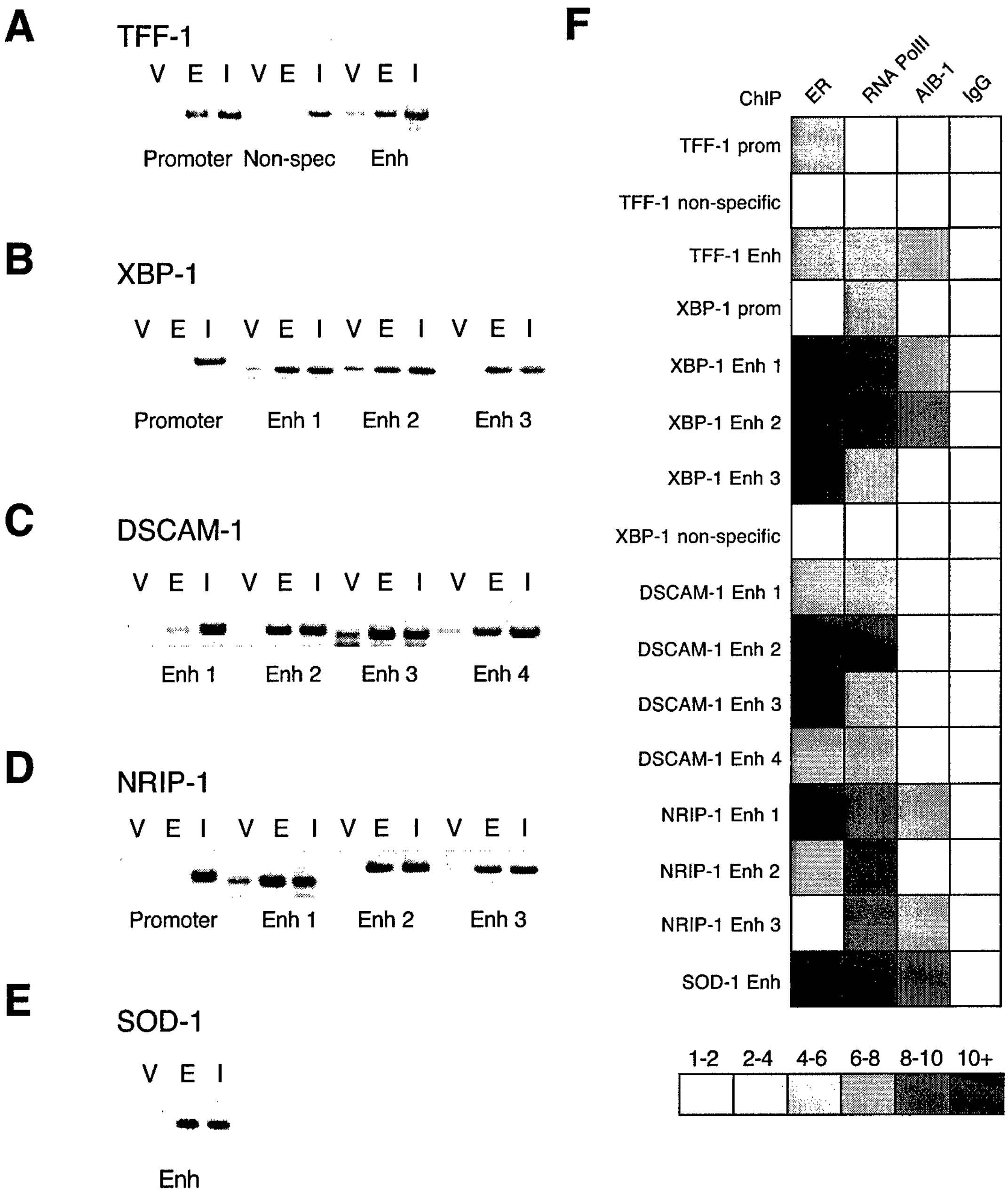

A
TFF-1
V E I V E I V E I
Promoter Non-spec Enh
B
XBP-1
V E I V E I V E I V E I
Promoter Enh 1 Enh 2 Enh 3
C
DSCAM-1
V E I V E I V E I V E I
Enh 1 Enh 2 Enh 3 Enh 4
D
NRIP-1
V E I V E I V E I V E I
Promoter Enh 1 Enh 2 Enh 3
E
SOD-1
V E I
Enh
F
ChIP
ER
RNA PolII
AIB-1
IgG
TFF-1 prom
TFF-1 non-specific
TFF-1 Enh
XBP-1 prom
XBP-1 Enh 1
XBP-1 Enh 2
XBP-1 Enh 3
XBP-1 non-specific
DSCAM-1 Enh 1
DSCAM-1 Enh 2
DSCAM-1 Enh 3
DSCAM-1 Enh 4
NRIP-1 Enh 1
NRIP-1 Enh 2
NRIP-1 Enh 3
SOD-1 Enh
1-2
2-4
4-6
6-8
8-10
10+

Table 1

Cluster number	Number of binding sites	Start	Stop	Closest regulated gene
1	1	10048850	10049271	
2	1	14600251	14600737	
3	1	15171656	15172273	
4	6	15467150	15738864	NRIP-1
5	1	17422343	17422868	
6	1	21532885	21533421	
7	1	29151881	29152882	
8	1	31821967	31822715	SOD-1
9	2	35021165	35027898	
10	1	35510057	35510719	
11	2	36480740	36487032	
12	1	38635468	38636783	
13	10	40363341	40675801	DSCAM-1
14	1	41911683	41912284	
15	1	42005946	42006169	PRDM15
16	2	42680784	42691725	TFF-1
17	1	42830736	42831350	
18	1	43564518	43565261	NDUFV3
19	2	45606461	45663897	
20	1	45790004	45790654	Col18A1
21	2	17159455	17194014	
22	1	19566341	19566809	
23	3	19822950	19945255	
24	3	27534171	27543908	XBP-1
25	1	28106122	28107112	AP1B1
26	1	28237489	28238464	
27	1	28519139	28520023	
28	2	30300284	30307434	PISD
29	2	37030766	37033295	
30	1	39371665	39372232	
31	1	41361325	41361720	Predicted
32	1	45100090	45100552	

Figure 3

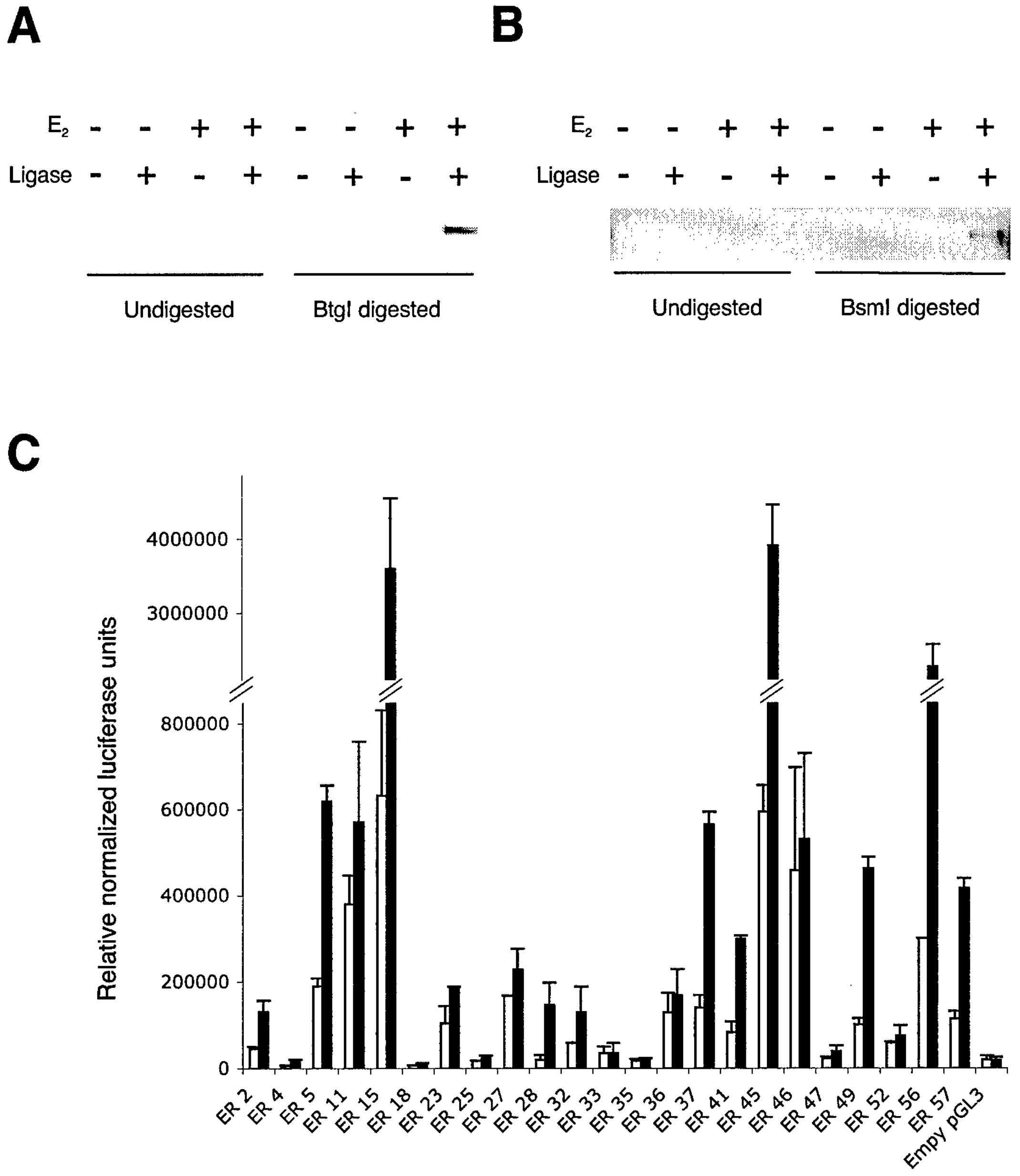

A
B
E_2 - - + + - - + +
Ligase - + - + - + - +
Undigested
BtgI digested
E_2 - - + + - - + +
Ligase - + - + - + - +
Undigested
BsmI digested
C
Relative normalized luciferase units
4000000
3000000
800000
600000
400000
200000
0
ER 2
ER 4
ER 5
ER 11
ER 15
ER 18
ER 23
ER 25
ER 27
ER 28
ER 32
ER 33
ER 35
ER 36
ER 37
ER 41
ER 45
ER 46
ER 47
ER 49
ER 52
ER 56
ER 57
Empy pGL3

Figure 4

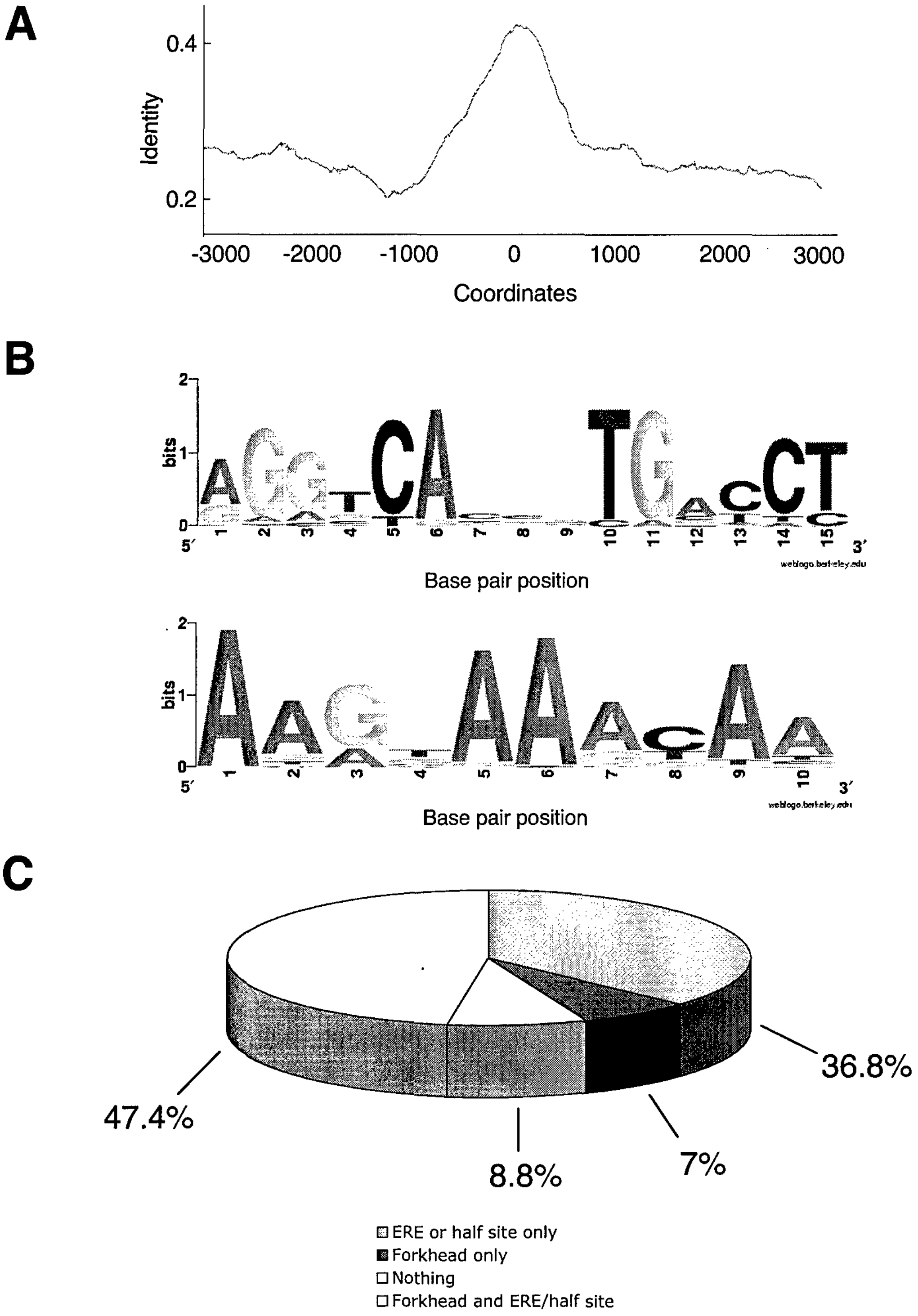

A
0.4
0.2
Identity
-3000
-2000
-1000
0
1000
2000
3000
Coordinates
B
2
1
0
bits
5′
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
3′
weblogo.berkeley.edu
Base pair position
2
1
0
bits
5′
1 2 3 4 5 6 7 8 9 10
3′
weblogo.berkeley.edu
Base pair position
C
36.8%
47.4%
8.8%
7%
▨ ERE or half site only
■ Forkhead only
□ Nothing
□ Forkhead and ERE/half site

Figure 5

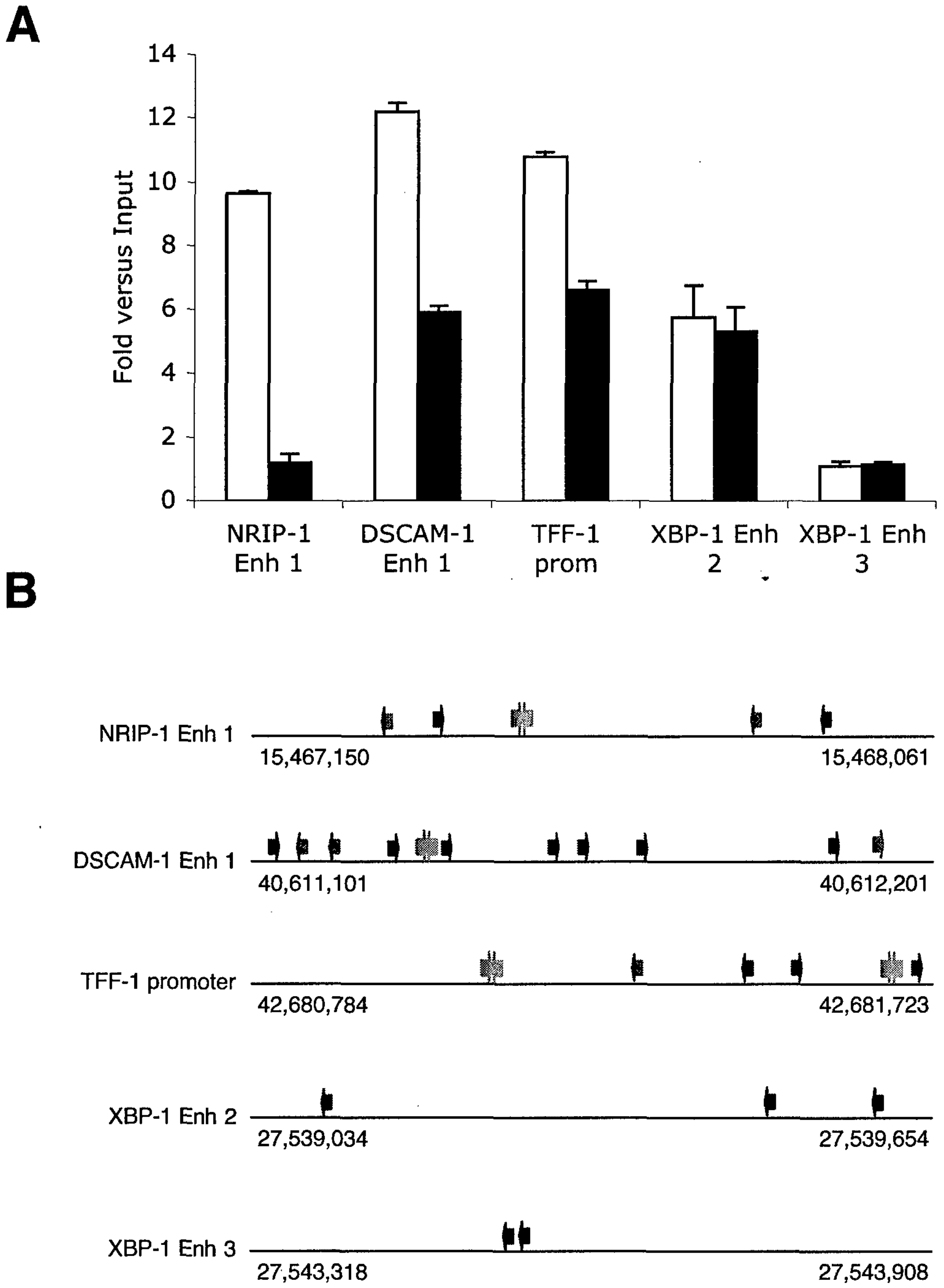
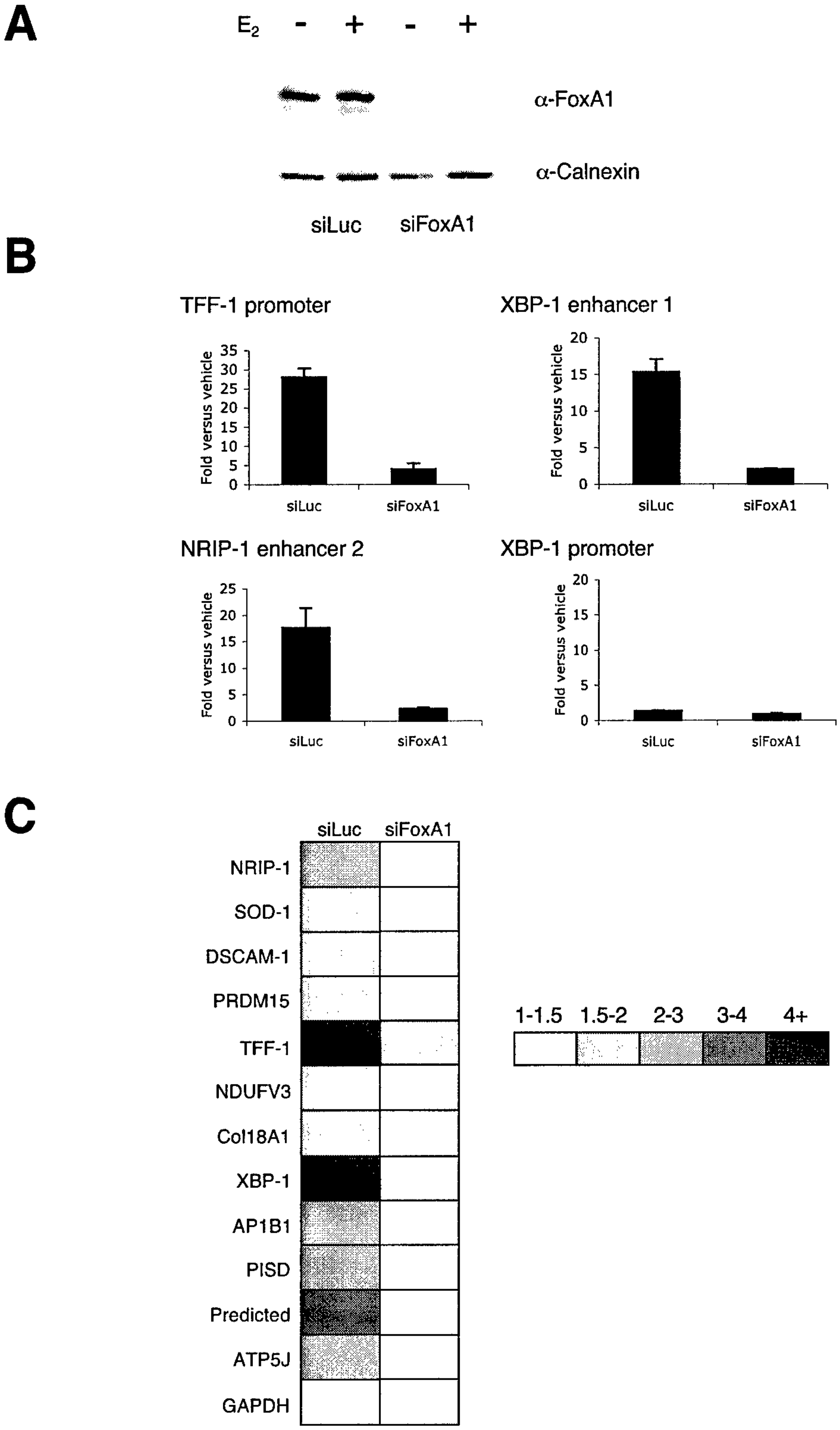

A
14
12
10
8
6
4
2
0
Fold versus Input
NRIP-1 Enh 1
DSCAM-1 Enh 1
TFF-1 prom
XBP-1 Enh 2
XBP-1 Enh 3
B
NRIP-1 Enh 1
15,467,150
15,468,061
DSCAM-1 Enh 1
40,611,101
40,612,201
TFF-1 promoter
42,680,784
42,681,723
XBP-1 Enh 2
27,539,034
27,539,654
XBP-1 Enh 3
27,543,318
27,543,908

Figure 6


A
E2 - + - +
α-FoxA1
α-Calnexin
siLuc siFoxA1
B
TFF-1 promoter
Fold versus vehicle
35
30
25
20
15
10
5
0
siLuc
siFoxA1
XBP-1 enhancer 1
Fold versus vehicle
20
15
10
5
0
siLuc
siFoxA1
NRIP-1 enhancer 2
Fold versus vehicle
25
20
15
10
5
0
siLuc
siFoxA1
XBP-1 promoter
Fold versus vehicle
20
15
10
5
0
siLuc
siFoxA1
C
siLuc siFoxA1
NRIP-1
SOD-1
DSCAM-1
PRDM15
TFF-1
NDUFV3
Col18A1
XBP-1
AP1B1
PISD
Predicted
ATP5J
GAPDH
1-1.5 1.5-2 2-3 3-4 4+

Genomic Mapping of RNA Polymerase II Reveals Sites of Co-Transcriptional Regulation in Human Cells

Alexander S. Brodsky[1]*, Clifford A. Meyer[2], Ian A. Swinburne[1], Giles Hall[3], Benjamin J. Keenan[1], Xiaole S. Liu[2], Edward A. Fox[3], and Pamela A. Silver[1]

[1] Department of Systems Biology, Harvard Medical School and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA

[2] Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, MA

[3] Department of Medicine, Harvard Medical School and Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA

*Person to whom correspondence should be addressed

Alexander S. Brodsky

Dana-Farber Cancer Institute

SM 922

44 Binney St.

Boston, MA 02115

Tel: 617 632 5105

alex_brodsky@dfci.harvard.edu

Abstract

Background

Transcription by RNA Polymerase II is regulated at many steps including initiation, promoter release, elongation and termination. Accumulation of RNA Polymerase II at particular locations across genes can be indicative of sites of regulation. RNA Polymerase II is thought to accumulate at the promoter and at sites of co-transcriptional alternative splicing where the rate of RNA synthesis slows.

Results

In order to further understand transcriptional regulation at a global level, we determined the distribution of RNA Polymerase II within regions of the human genome designated by the ENCODE project. Hypophosphorylated RNA Polymerase II localizes almost exclusively to 5' ends of genes. On the other hand, localization of total RNA Polymerase II reveals a variety of distinct landscapes across many genes with 74% of the observed enriched locations at exons. RNA Polymerase II accumulates at many annotated constitutively spliced exons, but is biased for alternatively spliced exons. Finally, RNA Polymerase II is also observed at locations not in gene regions.

Conclusions

Localizing RNA Polymerase II across many millions of base pairs in the human genome identifies novel sites of transcription and provides insights into the regulation of transcription elongation. These data indicate that RNA Polymerase II accumulates most often at exons during transcription. Thus, a major factor of transcription elongation control

in mammalian cells is the coordination of transcription and pre-mRNA processing to define exons.

Results and Discussion

Transcriptional and post-transcriptional regulation of gene expression intersect at RNA Polymerase II. The rate of RNA Polymerase II movement is altered by loading of transcription factors at the promoter, chromatin structure, pre-mRNA processing, elongation control and termination [1-3]. Thus, RNA Polymerase II accumulates at promoters as well as at different locations across a particular gene [4], but the general patterns across many different genes have yet to be explored. Numerous factors such as histones, post-translation modifying enzymes, and RNA binding proteins regulate these processes [1, 3]. One key determinant of transcription is the phosphorylation state of PolII's C-terminal domain (CTD) [5, 6] which becomes hyperphosphorylated during transcription elongation [4, 6-9]. Much of our understanding of transcription elongation comes from work in prokaryotes and yeast where most genes are intronless [1, 3]. Transcription and pre-mRNA processing are coordinated as the two processes affect the efficiency of each other [2, 10]. The spatial patterns of the different phosphorylation states of RNA Polymerase II across genes remains poorly understood in mammalian systems.

To explore the range of locations where RNA Polymerase II accumulates across the genome, we performed chromatin immunoprecipitation (ChIP) from HeLa S3 cells, and profiled the purified DNA using an oligonucleotide tiled microarray interrogating the ENCODE regions [11] covering 471 known genes. Two antibodies were utilized, 8WG16 and 4H8, which recognize the hypophosphorylated (PolIIa) or a phosphorylation independent state of the CTD of RNA Polymerase II (PolII), respectively. Thus, the 4H8 antibody is recognizing the total RNA Polymerase II population. Isolated DNA was amplified using a multiple displacement amplification (MDA) strategy (see Methods) [12].

To identify sites of enrichment, we used a non-parametric approach generalizing the Wilcoxon signed-rank test [13]. Signals across 1000 nucleotides were used to determine a p-value for each probe. Probes were filtered for uniqueness within the bandwidth. Probes with p-values below 10^{-4} were selected for further analysis because this threshold has a low false positive rate as determined by PCR analysis (Figure 1). With these parameters, the hypophosphorylated specific PolII antibody reveals 102 occupied sites whereas the phosphorylation independent antibody shows 550 sites (Table 1).

RNA Polymerase II has distinct landscapes across each gene. Figure 2 shows representative genes with PolII enrichments. PolIIa is highly enriched at transcription initiation sites. On the other hand, PolII shows gene-specific landscapes with the strongest enrichments at exons within actively transcribed loci. Active genes reveal lower p-values across the gene compared to intergenic or inactive genes (Compare Figure 1A and 1B) indicating a relative absence of RNA PolII from the nontranscribed regions. Some smaller genes with high exon density, such as SF1, reveal significant polymerase signal across almost the entire locus (Fig2A). Distinct accumulations are observed with significant p-values around exons for both SF1 and KIAA1932. In the KIAA1932 gene, RNA Pol II is enriched at a subset of constitutively and alternatively spliced exons (Figure 2C). For some genes, RNA Pol II is enriched at relatively few locations within the gene (Figure 2D).

An important question is to determine if the RNA Polymerase II sites are indicative of active transcription. We addressed this in multiple ways. First, microarray expression profiling of the mRNA with Affymetrix U133 Plus 2 chips confirms that many of the RNA Polymerase II associated genes are actively expressed in HeLa cells as seen in a plot of mRNA expression level vs. p-value in Figure 3. Genes with significant RNA Polymerase II enrichment are biased towards genes with higer mRNA levels. Figure 3 also shows that some genes have apparently high mRNA levels but no significant levels of PolII or PolIIa.

This could be due to very low transcription levels but high mRNA stability. Second, we measured RNA from the same HeLa cells on the ENCODE tiled arrays. We observe that 34% of the PolII sites overlap with RNA signal (compared to ~8% expected at random) while 50% of the PolII locations are within 1kb of some RNA signal (compared to 13% expected at random). Many sites where small pieces of RNA are synthesized such as small exons may be missed due to the spacing of the oligonucleotide probes and the imperfect nature of the probes. Third, many of the PolII and PolIIa sites overlap with annotated ESTs and mRNAs. 87% of the PolII and 88% of the PolIIa enriched locations overlap with EST regions, compared to 31% and 44% expected at random, respectively. Lastly, reverse transcriptase PCR checks of KIAA1932 and DKC1 indicate that these genes are being expressed (data not shown). These data suggest that RNA Polyermase II sites are biased towards region of active transcription and that determining sites of enrichment of RNA Polyermase II is an indicator of transcription.

Levels of PolII enrichment at internal exons can vary between genes. To examine whether these patterns are influenced by expression levels, two categories were created: I) Genes with multiple PolII enrichments at internal exons and II) Genes with PolII at one or zero internal exons. When compared to the mRNA levels, there is no significant difference between the two categories compared to mRNA levels suggesting that the number of PolII sites across the gene does not vary significantly with RNA levels. Genes with any PolII enrichment at internal exons are correlated with higher mRNA levels on the expression array. This is consistent with reports proposing to use PolII ChIP to monitor gene expression [14]. Therefore, the number of PolII sites at internal exons may reflect different levels of transcription elongation control and not just the sensitivity of the experiment.

Distinct from the hypophosphorylation specific antibody, the phosphorylation independent antibody reveals diverse enrichment locations for RNA PolII. In total, 74% of the identified PolII locations are near an annotated knownGene, Refseq or genscan exon as summarized in Table 1 (See Supplemental Table 2 for list of PolII genscan exon locations). Unlike PolIIa, PolII sites are distributed between the 5' and 3' ends of genes, with a slight bias towards terminal exons over initiating exons (Figure 4). This is likely reflecting RNA Polymerase II stalling during the coupled processes of transcription termination and 3'end processing [15]. For some genes, significant PolII signal is observed >1kb past the terminal exon, which might indicate transcription of the longer pre-mRNA before 3' end cleavage and polyadenylation [16]. Figure 5 shows two representative genes with significant RNA Polymerase II enrichment past the terminal exon.

Most of the hypophosphorylated PolII locations at internal exons also overlap a transcription initiation site as the internal exon in question is often the second exon in the gene. Only two enrichment sites overlap with an internal exon without also being near the first exon of a transcript. One of these is at a CpG island in the MCF2L gene and the other may be an alternative transcription initiation site as annotated in the HG17 assembly at the beginning of the ITGB4BP gene. To classify the remaining sites within introns or in intergenic regions, enrichment sites were compared to other gene databases. As summarized in Table 1, 4 PolIIa sites are in introns, but 3 of these are within resolution of annotated or predicted exons leaving only one location not overlapping an exon of some kind. There are 28 hypophosphorylated polymerase sites not in a Refseq gene region. After following a similar filtering approach, only 14 sites remain that are not near a putative exon. Thus, only 14% of of PolIIa enriched locations do not overlap with a known exon or actively transcribed region. Supplemental Table 2 lists PolIIa sites at predicted exons that are likely newly identified transcription initiation locations in HeLa cells. Figure 5 shows two examples of RNA Polymerase II and RNA signal at new sites of transcription. Based on the pattern of enrichments, it is probable that many of these predicted exons are real and are transcription initiation locations given the observed strong bias of the 8WG16 antibody for transcription initiation locations in well-annotated genes.

In order to determine the generality of these observations, all RNA PolII occupancy sites were compared to the known genes and Refseq databases, version HG15. PolIIa is highly enriched for the first exons around transcription initiation sites (Figure 4) representing 77 of 551 known genes in HG16 on the array (See Supplemental Tables 1A-1D for the entire lists.).

Elongation control is a common transcriptional regulation mechanism believed to be affecting a wide range of functional gene classes [1]. In particular, RNA Polymerase II pausing has been proposed to be associated with alternative splicing, [2]. To determine if there is a bias for alternative exons, we counted all the annotated alternatively spliced exons in the knownGene database and determined the distribution of PolII enrichment locations on them. PolII is enriched at 57% of the annotated alternatively spliced exons of the active genes compared to 37% of annotated actively transcribed constitutively expressed exons. We also examined the distribution of all PolII p-values on different types of exons. Each exon was mapped to the smallest p-value ChIP-enriched site that overlaps the exon. The cassette exons are found to be more significantly associated with smaller p-values compared to constitutively expressed exons according to the two-sample Kolmogorov-Smirnov test with a two sided p-value <0.0035.

One attractive hypothesis is that sites of exon enrichment may reflect weaker splice sites where PolII stalls during splice site recognition. Using two different empirical methods to estimate splice site strength, no significant differences are observed between the exons overlapping PolII and those that do not [17, 18]}. Alternatively, some of the annotated constitutively expressed exons may actually be subject to alternative splicing decisions. Kampa et al. suggest that the levels of alternative splicing are much higher than commonly believed and annotated in the human genome from their examination of expression on tiled arrays [19]. Consistent with these findings, RNA Polymerase II sites may be predicting which exons are being co-transcriptionally alternatively spliced.

To determine if there is any pattern for the 120 PolII enrichment sites that are in refSeq introns, we compared these sites to known gene, genscan, geneid, and sgpGene databases and find 31 within resolution of putative exons. Of the remaining 89, 57 are in genes with PolII enrichment sites that also overlap exons suggesting that they are actively transcribed genes. No clear intronic positional bias is observed.

In sum, we have identified new sites of RNA Polymerase II accumulation across hundreds of genes in mammalian cells. The large majority of RNA Polymerase II enriched locations are at actively transcribed exons with a bias towards annotated alternatively spliced exons. Many of the PolII sites at annotated constitutive expressed exons may be sites of alternative splicing. Whatever the eventual splicing decision, these observations suggest that events around exons slow transcription elongation. A recent study suggests that even general splicing factors may slow elongation [20]. Stalling of RNA Polymerase II near exons may function to slow RNA synthesis in order to wait for the competition of myriad splicing signals to be resolved in order to define the exon [21, 22]. These ChIP data identify where these states of RNA Polymerase II are localizing across the ENCODE regions.

Across genes, these data are consistent with the hypothesis of transcriptional pausing at particular locations. Alternatively, it is possible that RNA Polymerase II is rearranging during transcription such that the epitope is only accessible around exons. Thus, the conformation of RNA Polymerase II may be changing and not the transcription rate. Nonetheless, it is interesting that the majority of observable elongating RNA Polymerase II accumulates around exons suggesting that a major feature of transcription elongation control is coupling to pre-mRNA processing.

These observations differ from those observed in intronless genes typically found in prokaryotes and yeast where a more uniform PolII enrichment is observed across genes [16]. What appears to be conserved is PolII accumulation in coding regions compared to intronic regions. These data highlight the complexity and gene specific nature of transcription regulation not only at transcription initiation and termination locations but at specific exons. Together, these observations suggest that a major feature of transcription elongation control in mammalian cells is exon definition. Thus, these data provide new insights into the coordination of transcription and pre-mRNA processing in mammalian cells.

Materials and Methods

Chromatin Immunoprecipitation and DNA amplification. Chromatin immunoprecipitations were performed as described with the following modifications [23]. HeLa S3 cells were first crosslinked with DMA (Pierce) for 10 minutes, washed with PBS and then crosslinked with formaldehyde for 10 minutes. Cells were collected, lysed, and chromatin was sheared by sonication to an average length of 1 kb as determined after RNAse treatment of the samples on an agarose gel. Chromatin was prepared from four independently grown batches of cells and pooled to generate three replicate immunoprecipitatons and six input samples. Briefly, 8WG16 (Covance) and 4H8 (AbCam) antibodies were incubated with a 50:50 mix of Dynal protein A/G beads >16 hours at 4°C in PBS with 5 mg/ml BSA. After washing in PBS, beads with bound antibody were incubated with chromatin from approximately $2x10^7$ cells for > 16 hours at 4°C. Beads were washed 8 times with RIPA buffer (50 mM Hepes, pH 7.6, 1 mM EDTA, 0.7% DOC, 1% IGEPAL, 0.5 M LiCl) before DNA was eluted at 65°C in TE/1 % SDS. Crosslinks were reversed by incubating at 65°C for >12 hours followed by proteinase K treatment, phenol extraction and

RNAse treatment. Isolated DNA was then amplified isothermally using random nonamer primers and klenow polymerase (Invitrogen) for >4 hours yielding approximately 2 μg of DNA per IP. DNA was prepared and hybridized on Affymetrix ENCODE oligonucleotide tiled arrays using the fragmentation, hybridization, staining and scanning procedure described by Kennedy et al.[24]. Affymetrix ENCODE microarrays have an interrogating 25mer oligonucleotide probes tiled every 20 basepairs on average. A sample of chromatin was set aside before immunoprecipitation and used to represent the input DNA.

Tiled array analysis. Quantile normalization was used to make the distribution of probe intensities the same for all arrays [25]. In the case of the Affymetrix GTRANS software quantile normalization is used within treatment and control replicate sets. Nonparametric methods based on ranks were used to identify ChIP-enriched regions. These methods make mild assumptions about the data distributions and are insensitive to outlying observations. A p-value was calculated for every assay probe on the array. The set of probes used in the calculation of this p-value was defined by a bandwidth parameter b. All probes centered on the chromosome at positions less than b bases 5' or 3' of the given probe position are included in this set.

The Wilcoxon rank sum test [26], also known as the Mann-Whitney U test, is the basis of the p-value statistic computed by the Affymetrix GTRANS software. The control and treatment observation sets are, respectively, the sets of normalized control and normalized treatment intensities from all replicates and all probes within the bandwidth. The null hypothesis is that the treatment set mean is no larger than that of the control set.

To take into account probe to probe variability we used a generalization of the Wilcoxon signed-rank test for blocked data. All input and IP, normalized, sign(PM-MM)max(1,|PM-MM|) intensities interogating the same chromosomal location were assigned to the same block. Aligned observations were derived by subtracting the median, normalized intensity for a given block from each observation in that block. All aligned observations within the bandwidth were ranked. A statistic W was defined as the sum of the ranks of the aligned IP observations. A p-value was derived from W, based on the joint null distribution of the aligned input and IP ranks. The analyses are dependent on the assumption that probes are independent. Probes were mapped to the genomic coordinates to ensure that no probe mapped to more than one location in any 1000 bp window and no two probes map to the same genomic location.

RNA arrays. RNA samples were isolated from HeLa S3 cells and purified with trizol (Invitrogen) and RNAeasy (Qiagen). RNA was amplified and hybridized to Affymetrix U133 Plus 2 arrays using standard methods. Three biological replicates were quantile normalized. Gene expression was indicated by the median of PM-MM values over all probes. The hypothesis of difference in gene expression between groups of genes, based on median PM-MM was tested using the Wilcoxon rank sum statistic. For hybridization to the ENCODE tiled array, RNA was similarly isolated and double stranded cDNA was generated using Invitrogen Superscript cDNA synthesis kit. 1-1.5 ug of cDNA was hybridized to the tiled array. Three biological replicates were performed for each RNA array.

Genomic Annotation. Sites were determined to be near a genomic annotation if they were within the apparent 1000 bp resolution. Sites shorter than 1000 bp in length were scaled in size to include 1000 bp around the center of the site. Sites that were longer than 1000 bp used the data determined length for its resolution size. Databases were downloaded from the UCSC Golden Path Genome Browser and loaded into a local MySQL database. Exons were compared and classified as one or more of the following: Start, Terminal, Alternatively Spliced, Constitutive or Cassette. Because the arrays were designed using the HG15 assembly, the data were compared to this version of the human genome unless otherwise noted. The active gene list was defined as those with PolIIa at the first exon of the gene.

Real-time PCR PCR primer pairs were designed to amplify 100 bp fragments from selected genomic regions (Supplemental Table 8). Each real-time PCR reaction contained 50 nM primers, ~1ng of DNA and 1x ABI SYBR PCR Reaction Mix. A fluorescence value proportional to the initial quantity of target DNA was calculated by a log-linear regression analysis for each quadruplicate amplification curve [27]. We normalized this value to an input chromatin sample, then normalized this ratio to a reference gene, PAPT which is not expressed in HeLa cells, to calculate a relative enrichment value for the target $((\mathrm{Target}_{IP})/(\mathrm{Target}_{Inp}))/((\mathrm{PAPT}_{IP})/(\mathrm{PAPT}_{Input}))$.

Supplemental Data

1. PolIIa annotated to Refseq.
2. PolIIa annotated to known genes.
3. PolII annotated to Refseq.
4. PolII annotated to known genes.
5. PolII annotated to genscan exons.
6. knownGene and Refseq populations on the ENCODE array.
7. The PolIIa defined active gene list.
8. PCR primer list and annotation.

Figure Legends

Figure 1. Enrichment of Selected Genomic Regions in A) PolII ChIP B) PolIIa ChIP. Real-time PCR relative enrichment ratios for selected regions are found to be enriched more often with p-values below 10^{-4}. These regions include both intra- and intergenic locations as listed in Supplemental Table 8.

Figure 2. RNA Polymerase II shows a variety of gene specific enrichment patterns. Graphs plot –10log(p-value) mapped to chromosome position with the significant p-values greater than 40 indicated by the rectangle blocks below the graph. Values are plotted at every probe location. Flat lines indicate weak p-values and gaps indicate the absence of probes. The high density of probes across these genes suggest that the observed patterns are not due to probe bias. A scale bar is shown for each panel to reflect the different gene lengths displayed. RefSeq genes and known genes are annotated in green and blues, respectively, with thick bars representing exons and thin lines introns. Genes above the white bar are ordered 5'-3' while those below the white bar are 3'-5'. A. On the highly expressed SF1 gene, PolIIa localizes to the first exon only. PolII accumulates across the gene with a distinctive pattern. B. No significant signal is observed across the inactive NRXN2 locus that is nearby SF1 on chromosome 11. Graphs are plotted on the same scale as seen in A. C. The moderately expressed gene, KIAA1932, also reveals distinct accumulations across the gene. The red box highlights alternatively spliced exons. At the 3' end of the gene, some PolIIa signal is observed, probably indicative of the expression of a small gene antisense to the KIAA1932 gene. D. Another commonly observed pattern is exemplified by the EHD1 gene. Both polymerase antibodies recognize the first exon but no other significant signal is observed across the gene until the 3' end of the gene.

Figure 3. Different RNA Polymerase states show distinct exon biases. Pie charts representing the percentage of exons in each category at RNA Polymerase enrichment locations. These include exons from enrichment locations that include more than one exon. Hypophosphorylated polymerase is strongly biased towards transcription initiation locations. Most of the internal exons are second exons overlapping with first exons. The

phosphorylation independent antibody recognizes RNA Polymerase II at both transcription initiation and termination locations with a slight bias towards termination locations.

Figure 4. Low p-value PolII and PolIIa enrichments are biased towards higher mRNA levels. The plot depicts the observed intensity from Affymetrix U133 Plus 2 chips compared to different p-values of PolII (white) and PolIIa (gray). Some genes with no significant PolII enrichment show high levels of observed intensity.

Figure 5. RNA Polymerase II enrichment is not always within annotated gene boundaries as seen in views from the UCSC genome browser genome version HG16. PolIIa is in black and PolII is in blue with 4 rows for each representing the data at different p-values: $p<10^{-5}$, $p<10^{-4}$, $p<10^{-3}$, and $p<10^{-2}$ from top to bottom. RNA signal in red. Panels A and B show PolII extending beyond the 3' of the annotated gene. Panels C and D show RNA Polymerase II signal in putative intergenic regions with observed RNA signal also observed in the vicinity. Panel D covers chr11:285,000-290,000. These regions are conserved and are also near predicted genscan exons. These novel sites not in the gene regions were confirmed by PCR.

Acknowledgements

The authors would like to thank Pamela Hollasch, Maura Berkeley and the DFCI Affymetrix core for all their assistance. We thank Jason Carroll and Jessica Hurt for critical reading of the manuscript. We thank Adnan Derti for trying some splice site strength analysis. This work was funded by a NHGRI K22 career award, HG02488-01A1 (ASB), and a DOD grant DAMD17-02-0364 (PAS).

Table 1. Summary of RNA Polymerase II locations. The order indicates the flowchart of filtering through the different databases. Enrichment sites were first compared to the Refseq database. Sites that are not near exons were then divided into two categories: 1) Locations that are in Refseq introns. 2) Locations that are not in a Refseq gene are then compared to knownGene and predicted gene databases. For both RNA Polymerase II phosphorylation states, the large majority of sites are near an exon.

		Pol IIa	Pol II
Total Sites		102	550
Refseq total exons		70	289
Refseq first exons		63	75
Refseq terminal exons		2	91
Refseq internal exons		5	123
Refseq introns		4	120
	knownGene exon	0	5
	genscan exon	1	23
	geneid or sgpGene	0	3
	Active gene introns	2	57
	Inactive introns	1	32
No Refseq overlap		28	141
	knownGene total exons	5	38
	knownGene first exon	5	13
	knownGene terminal exon	0	4
	knownGene internal exon	0	21
	No Refseq or knownGene	23	90
	genscan exons	7	43
	geneid or sgpGene	2	6

References

1. Arndt KM, Kane CM: **Running with RNA polymerase: eukaryotic transcript elongation.** *Trends Genet* 2003, **19**(10):543-550.
2. Kornblihtt AR, de la Mata M, Fededa JP, Munoz MJ, Nogues G: **Multiple links between transcription and splicing.** *RNA* 2004, **10**(10):1489-1498.
3. Sims RJ, 3rd, Belotserkovskaya R, Reinberg D: **Elongation by RNA polymerase II: the short and long of it.** *Genes Dev* 2004, **18**(20):2437-2468.
4. Cheng C, Sharp PA: **RNA polymerase II accumulation in the promoter-proximal region of the dihydrofolate reductase and gamma-actin genes.** *Mol Cell Biol* 2003, **23**(6):1961-1967.
5. Dahmus ME: **Reversible phosphorylation of the C-terminal domain of RNA polymerase II.** *J Biol Chem* 1996, **271**(32):19009-19012.
6. Komarnitsky P, Cho EJ, Buratowski S: **Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription.** *Genes Dev* 2000, **14**(19):2452-2460.
7. Boehm AK, Saunders A, Werner J, Lis JT: **Transcription factor and polymerase recruitment, modification, and movement on dhsp70 in vivo in the minutes following heat shock.** *Mol Cell Biol* 2003, **23**(21):7628-7637.
8. Kim M, Ahn SH, Krogan NJ, Greenblatt JF, Buratowski S: **Transitions in RNA polymerase II elongation complexes at the 3' ends of genes.** *Embo J* 2004, **23**(2):354-364.
9. Ahn SH, Kim M, Buratowski S: **Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and 3' end processing.** *Mol Cell* 2004, **13**(1):67-76.
10. Hirose Y, Tacke R, Manley JL: **Phosphorylated RNA polymerase II stimulates pre-mRNA splicing.** *Genes Dev* 1999, **13**(10):1234-1239.
11. **The ENCODE (ENCyclopedia Of DNA Elements) Project.** *Science* 2004, **306**(5696):636-640.
12. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, Sun Z, Zong Q, Du Y, Du J *et al*: **Comprehensive human genome amplification using multiple displacement amplification.** *Proc Natl Acad Sci U S A* 2002, **99**(8):5261-5266.
13. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ *et al*: **Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs.** *Cell* 2004, **116**(4):499-509.
14. Sandoval J, Rodriguez JL, Tur G, Serviddio G, Pereda J, Boukaba A, Sastre J, Torres L, Franco L, Lopez-Rodas G: **RNAPol-ChIP: a novel application of chromatin immunoprecipitation to the analysis of real-time gene transcription.** *Nucleic Acids Res* 2004, **32**(11):e88.
15. Enriquez-Harris P, Levitt N, Briggs D, Proudfoot NJ: **A pause site for RNA polymerase II is associated with termination of transcription.** *Embo J* 1991, **10**(7):1833-1842.
16. Kim M, Krogan NJ, Vasiljeva L, Rando OJ, Nedea E, Greenblatt JF, Buratowski S: **The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II.** *Nature* 2004, **432**(7016):517-522.
17. Shapiro MB, Senapathy P: **RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression.** *Nucleic Acids Res* 1987, **15**(17):7155-7174.

18. Zhang MQ, Marr TG: **A weight array method for splicing signal analysis**. *Comput Appl Biosci* 1993, **9**(5):499-509.
19. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G *et al*: **Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22**. *Genome Res* 2004, **14**(3):331-342.
20. Ujvari A, Luse DS: **Newly Initiated RNA encounters a factor involved in splicing immediately upon emerging from within RNA polymerase II**. *J Biol Chem* 2004, **279**(48):49773-49779.
21. Roberts GC, Gooding C, Mak HY, Proudfoot NJ, Smith CW: **Co-transcriptional commitment to alternative splice site selection**. *Nucleic Acids Res* 1998, **26**(24):5568-5572.
22. Robson-Dixon ND, Garcia-Blanco MA: **MAZ elements alter transcription elongation and silencing of the fibroblast growth factor receptor 2 exon IIIb**. *J Biol Chem* 2004, **279**(28):29075-29084.
23. Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, Young RA, Dynlacht BD: **E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints**. *Genes Dev* 2002, **16**(2):245-256.
24. Kennedy GC, Matsuzaki H, Dong S, Liu WM, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J *et al*: **Large-scale genotyping of complex DNA**. *Nat Biotechnol* 2003, **21**(10):1233-1237.
25. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-193.
26. Hollander M, Wolfe DA: **Nonparametric Statistical Methods**, Second Edition edn. New York: John Wiley; 1999.
27. Ostermeier GC, Liu Z, Martins RP, Bharadwaj RR, Ellis J, Draghici S, Krawetz SA: **Nuclear matrix association of the human beta-globin locus utilizing a novel approach to quantitative real-time PCR**. *Nucleic Acids Res* 2003, **31**(12):3257-3266.

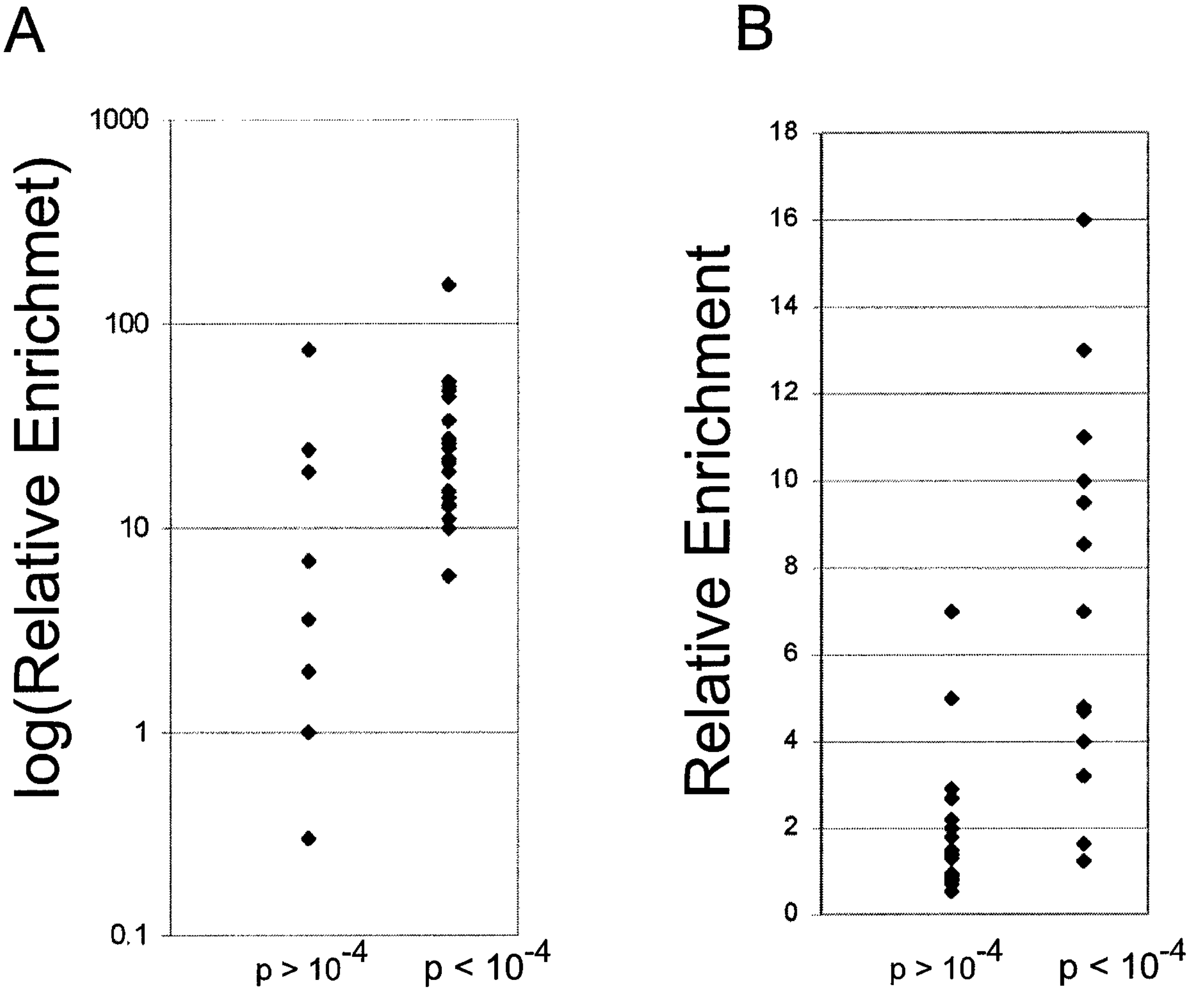

A
B
log(Relative Enrichmet)
1000
100
10
1
0.1
p > 10^{-4}
p < 10^{-4}
Relative Enrichment
18
16
14
12
10
8
6
4
2
0
p > 10^{-4}
p < 10^{-4}

Figure 1

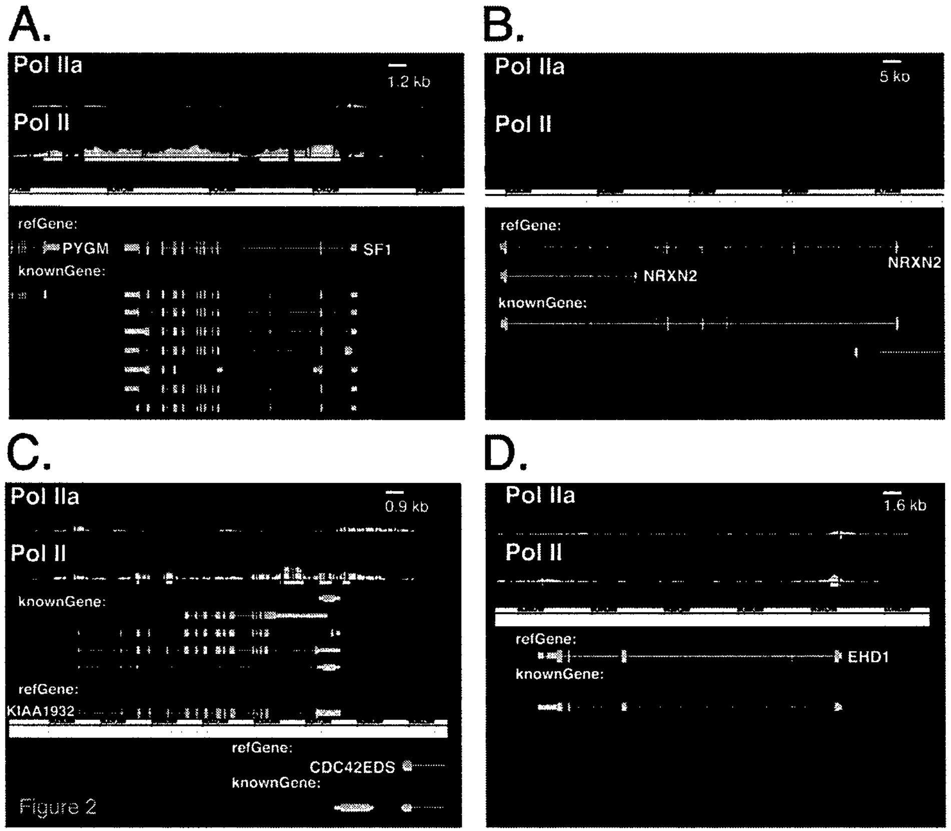

A.
Pol IIa
1.2 kb
Pol II
refGene:
PYGM
SF1
knownGene:
B.
Pol IIa
5 kb
Pol II
refGene:
NRXN2
NRXN2
knownGene:
C.
Pol IIa
0.9 kb
Pol II
knownGene:
refGene:
KIAA1932
refGene:
CDC42EDS
knownGene:
D.
Pol IIa
1.6 kb
Pol II
refGene:
EHD1
knownGene:

Figure 2

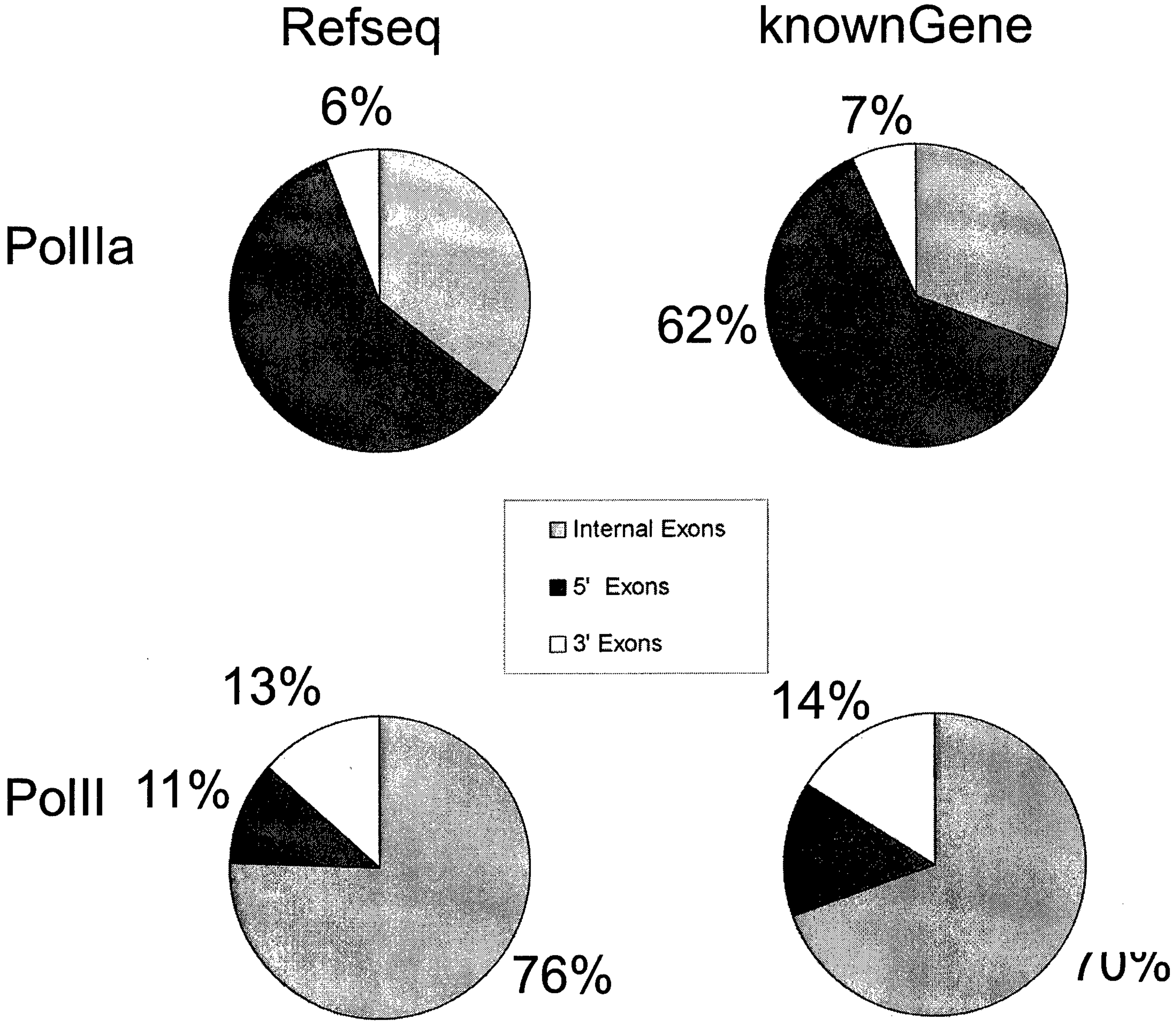

Refseq
knownGene
6%
7%
PolIIa
62%
▨ Internal Exons
■ 5' Exons
□ 3' Exons
13%
14%
PolII
11%
76%

Figure 3

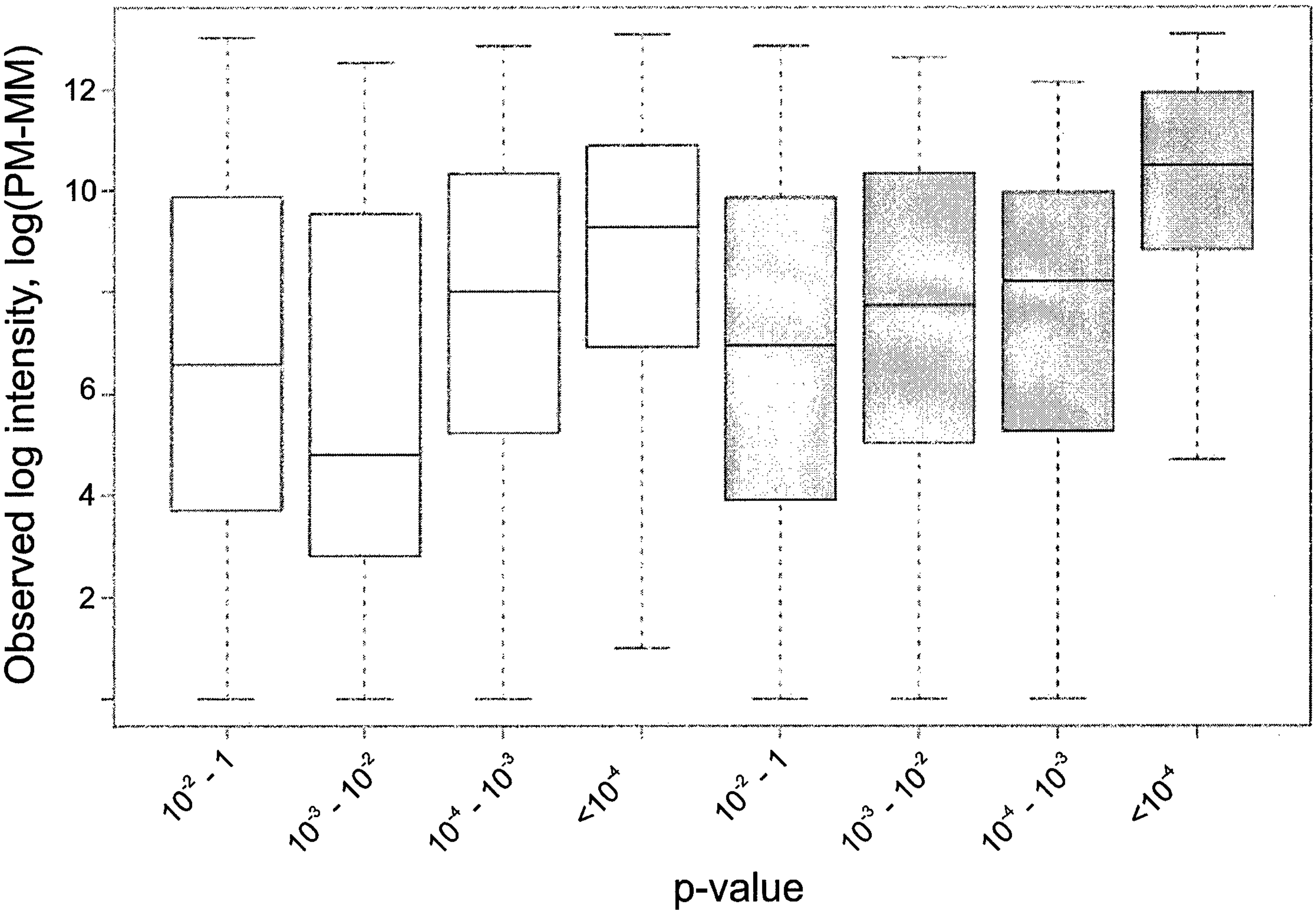

Observed log intensity, log(PM-MM)
12
10
6
4
2
10^{-2} - 1
10^{-3} - 10^{-2}
10^{-4} - 10^{-3}
<10^{-4}
10^{-2} - 1
10^{-3} - 10^{-2}
10^{-4} - 10^{-3}
<10^{-4}
p-value

Figure 4

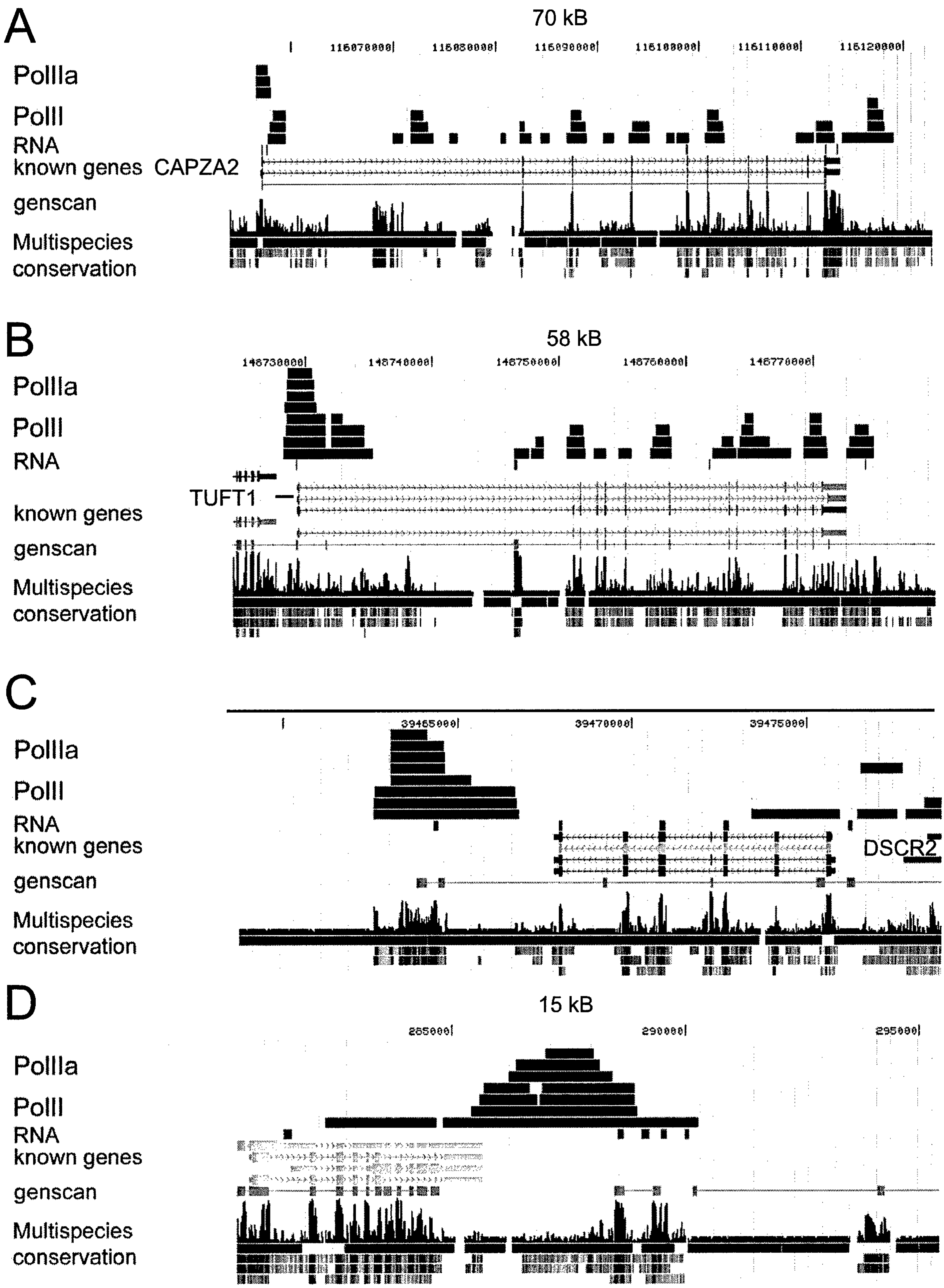

A
70 kB
116070000|
116080000|
116090000|
116100000|
116110000|
116120000|
PolIIa
PolII
RNA
known genes
CAPZA2
genscan
Multispecies
conservation
B
58 kB
148730000|
148740000|
148750000|
148760000|
148770000|
PolIIa
PolII
RNA
TUFT1
known genes
genscan
Multispecies
conservation
C
39465000|
39470000|
39475000|
PolIIa
PolII
RNA
known genes
DSCR2
genscan
Multispecies
conservation
D
15 kB
285000|
290000|
295000|
PolIIa
PolII
RNA
known genes
genscan
Multispecies
conservation

Figure 5

Additional files provided with this submission:

Additional file 1: SuppTable1.xls : 22KB
http://genomebiology.com/imedia/4799203816601869/sup1.xls

Additional file 2: SuppTable2.xls : 25KB
http://genomebiology.com/imedia/1706225883660187/sup2.xls

Additional file 3: SuppTable3.xls : 64KB
http://genomebiology.com/imedia/8661557366601788/sup3.xls

Additional file 4: SuppTable4.xls : 73KB
http://genomebiology.com/imedia/1973395791660178/sup4.xls

Additional file 5: SuppTable5.xls : 13KB
http://genomebiology.com/imedia/2100762506660179/sup5.xls

Additional file 6: SuppTable6.xls : 18KB
http://genomebiology.com/imedia/8303687166601791/sup6.xls

Additional file 7: SuppTable7.xls : 15KB
http://genomebiology.com/imedia/2014950166660179/sup7.xls

Additional file 8: SuppTable8.xls : 17KB
http://genomebiology.com/imedia/2124836802660188/sup8.xls

Analysis of RNA-protein interactions by flow cytometry

Alexander S Brodsky[1]*, Angus PR Johnston[2], Matt Trau[2] & Pamela A Silver[1]

Addresses
[1]Dana-Farber Cancer Institute and Department of Biological Chemistry and Molecular Pharmacology
Harvard Medical School
1 Jimmy Fund Way
SM 922
Boston
MA 02115
USA
Email: alex_brodsky@dfci.harvard.edu

[2]Centre for Nanotechnology and Biomaterials
Department of Chemistry
University of Queensland
St Lucia 4072
Brisbane
QLD
Australia

*To whom correspondence should be addressed

Current Opinion in Molecular Therapeutics 2003 5(3):235-240
© Current Drugs ISSN 1464-8431

Flow cytometry, in combination with advances in bead coding technologies, is maturing as a powerful high-throughput approach for analyzing molecular interactions. Applications of this technology include antibody assays and single nucleotide polymorphism mapping. This review describes the recent development of a microbead flow cytometric approach to analyze RNA-protein interactions and discusses emerging bead coding strategies that together will allow genome-wide identification of RNA-protein complexes. The microbead flow cytometric approach is flexible and provides new opportunities for functional genomic studies and small-molecule screening.

Keywords Flow cytometry, functional genomics, microbeads, microspheres, RNA-protein interactions

Introduction

The determination of RNA-protein regulatory networks is critical for understanding biological pathways. The role of RNA and RNA-protein interactions in regulating gene expression is becoming more appreciated with each new discovery. RNA-protein interactions are the backbone of many post-transcriptional processes, including mRNA stability, splicing, translation and localization. Determining which RNAs and proteins interact remains a challenging goal in the post-genomics era. Many human diseases such as fragile X [1] and HIV [2] are controlled by proteins interacting with RNAs. Proteins also form complexes with both large and small (eg, 7SK) non-coding RNAs [3,4] and microRNAs [5] to regulate gene expression. Understanding how RNA-protein interactions shape gene expression pathways on genome-wide levels remains unclear.

This review highlights recent advances in technologies to study RNA-protein interactions using genomic and high-throughput methods. In particular, we will focus on the use of microbeads to explore RNA-protein complexes by flow cytometry. These methods could evolve into diagnostic assays and high-throughput screens of pharmacological agents targeted to RNA-protein interactions. In this review, the assay will be introduced, and aspects of microbead technology important for the assay, such as microbead multiplexing and surface chemistry, will be discussed.

RNA-protein screening approaches

Many assays have been developed to examine nucleic acid protein interactions *in vitro*, including gel mobility shift, footprinting and filter binding. Hazbun and Fields performed a large-scale electrophoretic gel mobility shift assay (EMSA) to monitor DNA binding proteins from pools of glutathione-S-transferase (GST) yeast protein libraries [6]. However, similar to the other biochemical strategies, EMSA requires many manipulations making genome-wide screening labor intensive and complicated. In addition, these approaches require labeling of the RNA to monitor binding, making it difficult to pick a particular protein and determine the specifically binding RNAs.

A number of genetic methods have been developed for the analysis of RNA-protein interactions. One system that can screen for either RNA binding proteins or for RNA sequences is the three-hybrid assay [7]. However, long RNA sequences cannot be analyzed and certain sequences cause transcription termination [7]. A second genetic strategy is the Translational Repression Assay Procedure (TRAP) in yeast. This strategy works well with hairpin-containing RNA binding sites but has yet to be tested with a variety of RNA structures [8]. More recently, phage display methods have been developed with a model system to clone candidate proteins binding to a specific RNA sequence [9]. Genetic methods in mammalian cell lines, such as the Tat-fusion transcriptional activation system [10] and frameshifting assay [11], offer the ability to screen in the presence of potential binding partners. One drawback of these methods is that the complexes are forced to form in particular cellular compartments that may not be the native location. In addition, they often depend upon the generation of cDNA libraries that may be biased towards the most abundant messages and would also miss non-coding RNAs such as microRNAs.

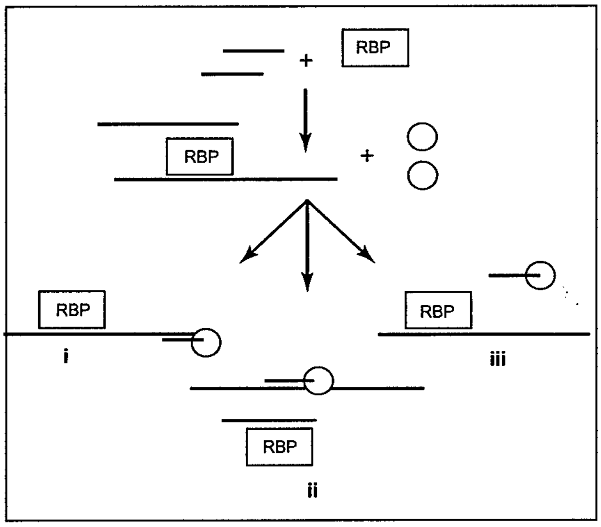
Recently, DNA chips have been used to identify RNAs bound to proteins [1,12,13•,14••]. This approach is promising for the investigation of RNA-protein interactions on a genome-wide scale. Typically, RNA-protein complexes are immunoprecipitated and the RNA is isolated and analyzed on DNA chips. Alternatively, protein can be prepared on beads and cell extract can be bound to the bead [15]. However, these approaches rely on the ability to preserve stable interactions during immunoprecipitation; many potentially weak interactions may be lost. In addition, RNA binding proteins typically have high non-specific binding constants leading to the isolation of a mixture of specific and non-specific 'bound' species, complicating the analysis. Other experiments such as systematic evolution of ligands by exponential amplification (SELEX) may be necessary to help determine the specifically binding RNAs [16].

The microbead assay

A new approach to RNA biochemistry uses flow cytometry and oligonucleotides attached to microbeads (Figure 1)

[17••]. A fluorescently labeled protein is bound to fluorescently labeled RNA before being challenged with oligonucleotides attached to microbeads. After hybridization, under conditions that do not disrupt the RNA-protein interaction, the microbeads are sorted and analyzed by flow cytometry. The presence or absence of RNA and protein signals provides binding interaction information. RNA-protein interactions can be specifically identified from complex mixtures while simultaneously characterizing binding properties, such as the dissociation constant (K_d). In addition, by probing with a high-density oligonucleotide library against RNAs of interest, the binding site could be determined.

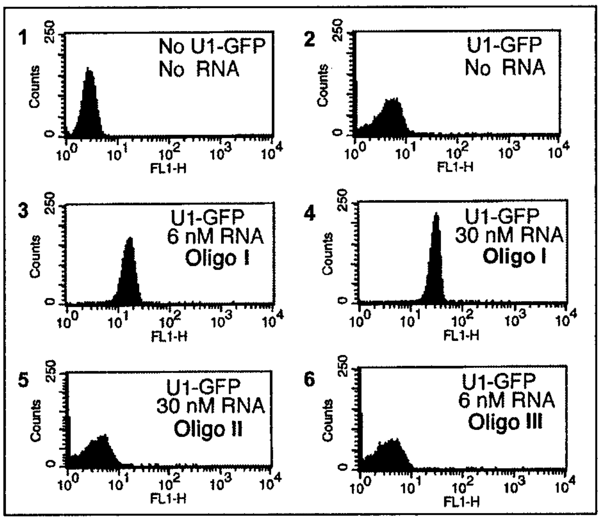
Figure 1. Schematic of the microbead assay.



A fluorescently labeled RNA binding protein (RBP)-RNA complex is formed and subsequently challenged with oligonucleotide beads. After reaching equilibrium, RNA and protein fluorescence on each microbead is determined by flow cytometry. The experiment can be performed with or without fluorescently labeled RNA. Three scenarios are possible: (i) RNA and protein fluorescence signal is observed indicating the bead is coupled to an oligonucleotide complementary to a RNA molecule that is binding the RBP; (ii) no protein fluorescence signal is observed but the oligonucleotide is hybridizing to the RNA. With labeled RNA, the RNA-oligonucleotide hybridization is detected. These oligonucleotides may be complementary to the RBP binding site and compete for RBP binding; and (iii) beads with neither protein nor RNA fluorescence suggesting that these oligonucleotides do not hybridize to the RNA. These sequences may be non-complementary to the RNA.

The RNA-protein microbead assay was developed with the U1 snRNP model system. Ul-green fluorescent protein (GFP) was purified and bound to a 150mer RNA; binding is indicated by GFP fluorescence on the bead population. (representative flow cytometry data are shown in Figure 2). RNA mutations, oligonucleotide mismatches and dissociation constants were measured to demonstrate the specificity of the assay. Single mismatch discrimination of short oligonucleotides was possible when the signal was monitored through the protein binding. Importantly, RNAs could be specifically detected in total RNA isolated from cells. The sensitivity is in the range of other common flow cytometry assays, since picomolar RNA concentrations could be detected. In this format, the assay is accessible to most molecular biology laboratories as it uses common reagents, and many facilities have access to flow cytometers.

Figure 2. Flow cytometry data showing GFP fluorescence on microbeads.



U1-GFP is bound to a 150mer RNA which includes a stem-loop binding site. Histograms show the number of beads at different GFP fluorescence intensities. In the presence of RNA and U1-GFP, the fluorescence intensity of the bead population increases and a more homogeneous bead population is observed, as shown by comparison of panels **2**, **3** and **4**. The different expected outcomes, as outlined in Figure 1, are shown: (i) oligonucleotide I is complementary to the RNA distant from the binding site; (ii) oligonucleotide II is complementary to the loop of the stem-loop and competes with U1-GFP binding; and (iii) oligonucleotide III is a non-complementary oligonucleotide. Oligonucleotides II and III show non-specific binding similar to the background, as shown by comparison of panels **2**, **5** and **6**.

The microbead assay is an equilibrium binding assay that offers some distinct advantages for the biochemical characterization of RNA-protein complexes. Firstly, protein binding to large RNAs can be examined. In fact, larger RNAs offer more hybridization targets for the antisense oligonucleotide probes. Using oligonucleotides targeting different regions of the RNA, binding can be monitored across the whole RNA molecule. Also, binding reactions could be performed in the presence of potential co-operative binding partners by using cell lysates or partially purified cell fractions. Since binding can be monitored at different locations across the RNA molecule, similar to footprinting assays, specific and non-specific sites may be differentiated. This may allow non-specific binding sites to be differentiated in genomic screens as the assay can monitor interactions from the picomolar to nearly micromolar dissociation constant range. Thus, weak interactions can be monitored and potentially discriminated from non-specific interactions.

The assay requires fluorescent labeling of the RNA and/or protein. A number of strategies have been, and are continuing to be developed to label proteins with minimal disruption to their structure and function. The first generation of the RNA-protein bead assay uses a GFP fusion. Other strategies include the use of antibodies, where a fluorescently labeled antibody against a protein is used to monitor binding [18]. Screening in yeast has found that many proteins are functional when either their N- or C-terminal are tagged [19]. Advances in *in vitro* translation may also allow site-specific labeling at the beginning or end of the protein as seen in both Promega and Roche systems. These strategies will increase the probability of obtaining functionally active, fluorescently labeled protein.

Flow cytometry

Flow cytometry is a powerful, sensitive and quantitative technology used to measure molecular interactions. Flow cytometry has been successfully applied to examine various protein-protein, protein-DNA and DNA-DNA interactions [20,21]. As it is fluorescence based, it can also be adapted to monitor real-time kinetics and rapid quench studies. Very high sensitivity can be obtained with 10^5 to 10^6 particles/ml and target concentrations in the 10 to 100 pM range, well below the K_d of most RNA-protein interactions. Since flow cytometry can focus on just the signal on the microbeads and not the unbound molecules in solution, typically, no washing is required, saving significant effort. Also, recent advances in coding microbeads are bringing the power of multiplexing dozens of samples simultaneously to these assays. With the ability to use automatic sample loaders running at two to three samples/min, high-throughput plate reading is now feasible. Recent reviews highlight the latest technical advances in flow cytometry, allowing for high throughput and sensitivity [20-22].

Microbeads versus microarrays and hybridization

Binding to microbeads instead of microarrays offers a number of potentially significant advantages. Microbeads have proven to be useful for sensitive and rapid bioanalytical assays. Companies such as Luminex Corp, Lynx Therapeutics and Quantum Dot Corp have taken advantage of these properties to devise high-throughput approaches to immunoassays [23], sequencing [24,25] and single nucleotide polymorphism mapping [26,27••]. Importantly, microbead assays are typically cost-effective, fast and require minimal sample quantities.

An important advantage of microbead for the RNA-protein binding assay is the ability to perform binding on a surface that more closely resembles solution conditions. Hybridization on large planar surfaces is limited by mass transport. On the other hand, microbeads offer better diffusion characteristics, leading to significantly improved hybridization kinetics and thermodynamics [28•,29•].

The basic approach of the microbeads assay is also applicable to microarrays. However, non-specific hybridization at physiological conditions is a requirement of the assay. Due to the demanding hybridization requirements and the relative ease in synthesizing oligonucleotides with long linkers to readily available microbeads, the microbead approach offers a simple alternative to microarrays.

Encoding strategies

Bead libraries

Most high-throughput bead-based libraries use the optical properties of the support as the library code. The exception to this is the approach from Lynx Therapeutics, who utilize non-encoded support beads, and a series of molecular markers and identifiers [24]. Optical encoding of supports falls into two broad categories. The first (Luminex Corp, Quantum-dot Corp, Illumina and Nanoplex) is based on separately coding each bead and separately synthesizing the target DNA sequence (or other analyte such as RNA or peptide), then attaching each target to a coded bead. The alternative technique is to directly synthesize the target molecule on a coded bead in a combinatorial manner and track every synthetic step each individual bead experiences [30••].

Separate encoding

Methods that use the separate encoding strategy employ a similar strategy to encode the beads. In each case, fluorochromes, fluorescent dyes (Luminex [31] and Illumina [32]) or fluorescent nanocrystals (Quantum-dot Corp [27••,33]) are incorporated into polystyrene beads by swelling the polystyrene in a solvent and absorbing dyes or nanocrystals into the particles. The bead is then placed in a different solvent to shrink the bead, trapping the fluorochrome in the bead. The code is formed by varying the concentration and the combination of fluorochromes present in each bead. The code can be read either by a flow cytometer (Luminex Corp, Quantum-dot Corp) or by optic fiber array (Illumina). DNA sequences are synthesized remotely (either separately in an automated DNA synthesizer or *in vivo*) and attached to the beads using standard ethylenediamine carbodiimide (EDC) coupling chemistry [34]. The separate encoding techniques are useful for small libraries since it is easy to separately synthesize hundreds of different beads and hundreds of target molecules (Table 1). However, there are limitations for larger libraries [35]. To synthesize a library of 100,000 compounds requires 100,000 separate coded beads and 100,000 separately synthesized DNA sequences combined in 100,000 coupling reactions (Figure 3). Automation of this process is possible; however, the size of the library is still limited by the number of coded beads that can be formed.

Nanoplex uses metallic rods (instead of spherical particles) with bands of material with different refractive index to form coding system, which is similar to traditional barcodes but on a microscopic scale [36]. The difference in refractive index is achieved by incorporating different metals into the rods as they are synthesized. Similar to the other separate encoding strategies, library size is limited by the number of separate reaction vessels required to synthesize the coded support and the analyte. However, unlike the fluorescent coding approach, the barcode can be incorporated over a large number of steps, so the coding system can code for many more sequences than it would be possible to synthesize in the library. At this time, there is no automated, high-throughput method of reading these barcodes.

Combinatorial encoding/synthesizing

In the combinatorial method, a set of optically diverse, but distinguishable particles are used as the support for synthesizing the target DNA (or other target molecules) [37]. The optically diverse set of particles are synthesized using a combinatorial process where beads are split into a number

Table 1. Comparison of microbead coding strategies.

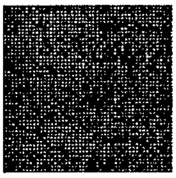
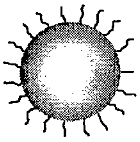
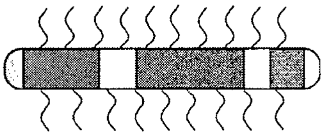
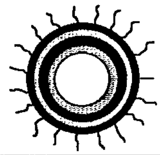
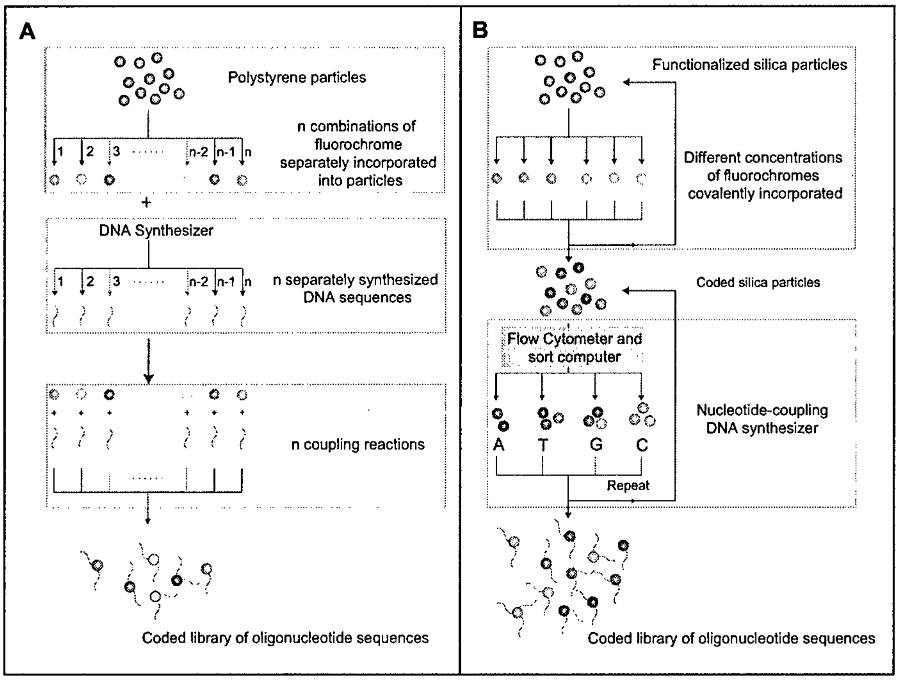
	Encoding method	Decoding method	Library size
 DNA microarray	Positional encoding. Probes are immobilized in spatially resolved sites on a two-dimensional support.	Via position in array.	< 10^6 probes [42].
 Probe attached to bead	Non-permanently stained polymer beads with up to four fluorochromes.	Flow cytometry [31,33], optical fiber arrays [30••], digital imaging.	100 to 270,000 probes.
 Probe attached to rod	Layered metallic strips on rod shaped particles.	Microscopic imaging (not automated).	> 100 probes [36] (can potentially code 10^{13} but library size is limited by decoding rate and library synthesis).
 Probe synthesized on bead	The unique optical signature of each multi-fluorescent support bead is tracked by a flow cytometer during the combinatorial synthesis of the probe.	The optical signature is analyzed by flow cytometry and the reaction history of the bead is determined by recalling data stored by the flow cytometer software during probe synthesis.	>10^8 probes [30••].

Figure 3. Comparison of encoding techniques.


A
Polystyrene particles
n combinations of fluorochrome separately incorporated into particles
1
2
3
......
n-2
n-1
n
+
DNA Synthesizer
1
2
3
......
n-2
n-1
n
n separately synthesized DNA sequences
n coupling reactions
......
Coded library of oligonucleotide sequences
B
Functionalized silica particles
Different concentrations of fluorochromes covalently incorporated
Coded silica particles
Flow Cytometer and sort computer
Nucleotide-coupling DNA synthesizer
A
T
G
C
Repeat
Coded library of oligonucleotide sequences

(**A**) Separate encoding strategy where beads are individually coded in separate reaction vessels and the oligonucleotides are individually synthesized remotely. The oligonucleotide is coupled to the bead using standard EDC chemistry.
(**B**) Combinatorial encoding strategy where silica particles are coded using a split and mix process with varying concentrations of dyes. Using a customized flow cytometer, the particles are sorted into four reaction vessels (one for each base) according to predetermined parameters. The process is repeated until oligonucleotides of the required length are synthesized.

of reaction vessels and varying concentrations of fluorophores such as organic fluorescent dyes or nanoparticles, are covalently incorporated into the beads. The beads are then mixed together and the process is repeated for each subsequent dye; thus, it is not necessary to synthesize each coded bead individually [30••]. Using this method with six fluorophores and eight levels of intensity for each dye, a library of over 250,000 signatures can be constructed. Still using only six fluorophores, but with 16 levels of intensity, a library of over 16 million sequences can be generated (Table 1).

Using a flow cytometer and custom designed electronics, beads can be analyzed and sorted according to the particular optical signatures [30••]. Each bead has a predetermined sequence that is uploaded to the modified flow cytometer and sort decisions are made according to the sequences that are required for the particular library. The flow cytometer can sort into four directions, with each direction corresponding to a different nucleoside. After each 'sort' the nucleosides are coupled to the corresponding beads and once coupling is complete, the beads are mixed together and the process is repeated until the oligonucleotide sequences of the required length are synthesized. At the end of the process, beads with a known optical signature are synthesized with each unique signature corresponding to a different oligonucleotide sequence (Figure 3).

Synthesizing libraries in this way requires beads that can withstand the relatively harsh conditions of DNA synthesis. Polystyrene beads are typically not suitable for this process because they swell and leach dye during the synthesis procedure, thus catastrophically altering the optical signature. Therefore, specially synthesized silica particles are required [30••].

Surface chemistry

The surface chemistry of the beads plays an important role in the assay. Non-specific binding of proteins to the beads is a larger problem than the non-specific binding of oligonucleotides to the beads, as generally, most surfaces with a large negative charge (eg, silica and polystyrene surfaces) have relatively low non-specific binding of oligonucleotides (due to repulsion of the negatively charged phosphate backbone). However, as proteins have positively and negatively charged regions, it is necessary to have a surface which has little or no surface charge to minimize non-specific electrostatic binding and molecule adsorption [38•]. Coating the surfaces with hydrophobic chains (such as alkyl chains) is also not ideal, since many proteins have hydrophobic regions that will also non-specifically bind to the beads [39]. One solution is to add a large excess of a protein (eg, inexpensive and abundant proteins such as bovine serum albumin) that non-specifically bind to the surface of the beads, limiting the non-specific binding of the fluorescently labeled protein. However, there is a limit to the effectiveness of this procedure and it is desirable to have a 'biologically silent' surface that limits the non-specific binding of the proteins.

Much of the surface chemistry developed for protein-chips can be applied to bead surfaces. Polyethylene glycol and oligoethylene glycol surfaces have been used to minimize the non-specific binding of proteins [39] to silica substrates.

Surface density of the probes also plays an important role in the assay. Clearly, the higher the number of probes on the beads, the higher the resultant signal; however, overloading the surface introduces problems. It is possible to load in excess of 100 million target sequences on a single bead, but at this very high surface density, steric hindrance can affect the hybridization of target DNA to the beads. In addition, false hybridization events may occur, where one target DNA strand hybridizes to multiple probe strands on the bead [40]. Similar findings have been observed on arrays [41].

High-throughput screening and genomics

This review describes the recent development of a versatile flow cytometry approach to examine RNA-protein interactions. The emerging bead-coding and surface chemistry technologies, in combination with novel assays such as the microbead RNA-protein assay will lead to new small molecule and genomic screens. Due to the versatility and flexibility of flow cytometry and the RNA-protein assay, many variations are possible, including defining the binding spectrum of a particular RNA-binding protein, screening a protein library for binding to a specific RNA, or discovering small molecules that inhibit an RNA-protein interaction. With the increasing understanding of the importance of RNA-protein interactions in human disease and development, the contribution of these promising technologies is expected to be significant.

Acknowledgements

Alexander S Brodsky is supported by the Claudia Adams Barr Young Investigator program and an NHGRI Genome Scholar award.

References

•• of outstanding interest

• of special interest

1. Brown V, Jin P, Ceman S, Darnell JC, O'Donnell WT, Tenenbaum SA, Jin X, Feng Y, Wilkinson KD, Keene JD, Darnell RB *et al*: **Microarray identification of FMRP-associated brain mRNAs and altered mRNA translational profiles in fragile X syndrome.** *Cell* (2001) **107**(4):477-487.

2. Frankel AD, Young JA: **HIV-1: Fifteen proteins and an RNA.** *Annu Rev Biochem* (1998) **67**:1-25.

3. Nguyen VT, Kiss T, Michels AA, Bensaude O: **7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes.** *Nature* (2001) **414**(6861):322-325.

4. Yang Z, Zhu Q, Luo K, Zhou Q: **The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription.** *Nature* (2001) **414**(6861):317-322.

5. Dostie J, Mourelatos Z, Yang M, Sharma A, Dreyfuss G: **Numerous microRNPs in neuronal cells containing novel microRNAs.** *RNA* (2003) **9**(2):180-186.

6. Hazbun TR, Fields S: **A genome-wide screen for site-specific DNA-binding proteins.** *Mol Cell Proteomics* (2002) **1**(7):538-543.

7. Zhang B, Kraemer B, SenGupta D, Fields S, Wickens M: **Yeast three-hybrid system to detect and analyze RNA-protein interactions.** *Methods Enzymol* (2000) **318**:399-419.

8. Paraskeva E, Hentze MW: **Translational repression assay procedure: A method to study RNA-protein interactions in yeast.** *Methods Enzymol* (2000) **318**:374-384.

9. Danner S, Belasco JG: **T7 phage display: A novel genetic selection system for cloning RNA-binding proteins from cDNA libraries.** *Proc Natl Acad Sci USA* (2001) **98**(23):12954-12959.

10. Landt SG, Tan R, Frankel AD: **Screening RNA-binding libraries using Tat-fusion system in mammalian cells.** *Methods Enzymol* (2000) **318**:350-363.

11. Kollmus H, Hauser H: **Frameshifting assay to characterize RNA-protein interactions in eukaryotic cells.** *Methods Enzymol* (2000) **318**:363-374.

12. Andrews LG, Keene JD: **Identification of specific protein-RNA target sites using libraries of natural sequences.** *Methods Mol Biol* (1999) **118**:233-244.

13. Tenenbaum SA, Carson CC, Lager PJ, Keene JD: **Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays.** *Proc Natl Acad Sci USA* (2000) **97**(26):14085-14090.

• *The first report of microarrays to analyze co-immunoprecipitating mRNAs.*

14. Hieronymus H, Silver PA: **Genome-wide analysis of RNA-protein interactions illustrates specificity of the mRNA export machinery.** *Nat Genet* (2003) **33**(2):155-161.

•• *A thorough microarray analysis of co-immunoprecipitating bound mRNAs.*

15. Trifillis P, Day N, Kiledjian M: **Finding the right RNA: Identification of cellular mRNA substrates for RNA-binding proteins.** *RNA* (1999) **5**(8):1071-1082.

16. Darnell JC, Jensen KB, Jin P, Brown V, Warren ST, Darnell RB: **Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function.** *Cell* (2001) **107**(4):489-499.

17. Brodsky AS, Silver PA: **A microbead-based system for identifying and characterizing RNA-protein interactions by flow cytometry.** *Mol Cell Proteomics* (2002) **1**(12):922-929.

•• *Description of the novel microbead assay.*

18. Yang X, Li X, Prow TW, Reece LM, Bassett SE, Luxon BA, Herzog NK, Aronson J, Shope RE, Leary JF, Gorenstein DG: **Immunofluorescence assay and flow-cytometry selection of bead-bound aptamers.** *Nucleic Acids Res* (2003) **31**(10):e54.

19. Phizicky E, Bastiaens PI, Zhu H, Snyder M, Fields S: **Protein analysis on a proteomic scale.** *Nature* (2003) **422**(6928):208-215.

20. Iannone MA: **Microsphere-based molecular cytometry.** *Clin Lab Med* (2001) **21**(4):731-742.

21. Sklar LA, Edwards BS, Graves SW, Nolan JP, Prossnitz ER: **Flow cytometric analysis of ligand-receptor interactions and molecular assemblies.** *Annu Rev Biophys Biomol Struct* (2002) **31**:97-119.

22. Nolan JP, Sklar LA: **The emergence of flow cytometry for sensitive, real-time measurements of molecular interactions.** *Nat Biotechnol* (1998) **16**(7):633-638.

23. Vignali DA: **Multiplexed particle-based flow cytometric assays.** *J Immunol Methods* (2000) **243**(1-2):243-255.

24. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R *et al*: **Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays.** *Nat Biotechnol* (2000) **18**(6):630-634.

25. Brenner S, Williams SR, Vermaas EH, Storck T, Moon K, McCollum C, Mao JI, Luo S, Kirchner JJ, Eletr S, DuBridge RB *et al*: ***In vitro*** **cloning of complex mixtures of DNA on microbeads: Physical separation of differentially expressed cDNAs.** *Proc Natl Acad Sci USA* (2000) **97**(4):1665-1670.

26. Taylor JD, Briley D, Nguyen Q, Long K, Iannone MA, Li MS, Ye F, Afshari A, Lai E, Wagner M, Chen J *et al*: **Flow cytometric platform for high-throughput single nucleotide polymorphism analysis.** *Biotechniques* (2001) **30**(3):661-666.

27. Xu H, Sha MY, Wong EY, Uphoff J, Xu Y, Treadway JA, Truong A, O'Brien E, Asquith S, Stubbins M, Spurr NK *et al*: **Multiplexed SNP genotyping using the Qbead system: A quantum dot-encoded microsphere-based assay.** *Nucleic Acids Res* (2003) **31**(8):e43.

•• *Description of the use and advantages of a quantum dot-based optical code.*

28. Henry MR, Wilkins Stevens P, Sun J, Kelso DM: **Real-time measurements of DNA hybridization on microparticles with fluorescence resonance energy transfer.** *Anal Biochem* (1999) **276**(2):204-214.

• *Describes experiments comparing microbead and solution phase hybridization kinetics.*

29. Wilkins Stevens P, Henry MR, Kelso DM: **DNA hybridization on microparticles: Determining capture-probe density and equilibrium dissociation constants.** *Nucleic Acids Res* (1999) **27**(7):1719-1727.

• *Describes experiments comparing microbead and solution phase hybridization kinetics.*

30. Battersby BJ, Lawrie GA, Johnston AP, Trau M: **Optical barcoding of colloidal suspensions: Applications in genomics, proteomics and drug discovery.** *Chem Commun (Camb)* (2002) **14**:1435-1441.

•• *Description of the combinatorial coding technique.*

31. Fulton RJ, McDade RL, Smith PL, Kienker LJ, Kettman JR Jr: **Advanced multiplexed analysis with the FlowMetrix system.** *Clin Chem* (1997) **43**(9):1749-1756.

32. Walt DR: **Techview: Molecular biology. Bead-based fiber-optic arrays.** *Science* (2000) **287**(5452):451-452.

33. Han M, Gao X, Su JZ, Nie S: **Quantum-dot-tagged microbeads for multiplexed optical coding of biomolecules.** *Nat Biotechnol* (2001) **19**(7):631-635.

34. Walsh MK, Wang X, Weimer BC: **Optimizing the immobilization of single-stranded DNA onto glass beads.** *J Biochem Biophys Methods* (2001) **47**(3):221-231.

35. Battersby BJ, Grondal L, Lawrie GA, Trau M: **Colloids for encoding chemical libraries: Applications in biological screening.** In: *Colloids and Colloidal Assemblies.* Caruso F (Ed), Wiley, New York, NY, USA (2002).

36. Nicewarner-Pena SR, Freeman RG, Reiss BD, He L, Pena DJ, Walton ID, Cromer R, Keating CD, Natan MJ: **Submicrometer metallic barcodes.** *Science* (2001) **294**(5540):137-141.

37. Trau M, Battersby BJ: **Novel colloidal materials for high-throughput screening applications in drug discovery and genomics.** *Advanced Materials* (2001) **13**(12-13):975-979.

38. Song J, Choi H, Moeser G, Laibinis P: **Preparation of protein- and cell-resistant surfaces on SiO_2 surfaces by polyther-grafted poly(acrylic acid) thin films.** *Polymer Preprints* (2003) **44**:189-190.

• *Describes the advantages of using ethylene glycol surfaces to minimize non-specific binding.*

39. Lee SW, Laibinis PE: **Protein-resistant coatings for glass and metal oxide surfaces derived from oligo(ethylene glycol)-terminated alkytrichlorosilanes.** *Biomaterials* (1998) **19**(18):1669-1675.

40. Guo Z, Gatterman MS, Hood L, Hansen JA, Petersdorf EW: **Oligonucleotide arrays for high-throughput SNPs detection in the MHC class I genes: HLA-B as a model system.** *Genome Res* (2002) **12**(3):447-457.

41. Peterson AW, Heaton RJ, Georgiadis RM: **The effect of surface probe density on DNA hybridization.** *Nucleic Acids Res* (2001) **29**(24):5163-5168.

42. Hacia JG: **Resequencing and mutational analysis using oligonucleotide microarrays.** *Nat Genet* (1999) **21**(1 Suppl):42-47.

A Microbead-based System for Identifying and Characterizing RNA-Protein Interactions by Flow Cytometry*

Alexander S. Brodsky‡§ and Pamela A. Silver

We present a high throughput, versatile approach to identify RNA-protein interactions and to determine nucleotides important for specific protein binding. In this approach, oligonucleotides are coupled to microbeads and hybridized to RNA-protein complexes. The presence or absence of RNA and/or protein fluorescence indicates the formation of an oligo-RNA-protein complex on each bead. The observed fluorescence is specific for both the hybridization and the RNA-protein interaction. We find that the method can discriminate noncomplementary and mismatch sequences. The observed fluorescence reflects the affinity and specificity of the RNA-protein interaction. In addition, the fluorescence patterns footprint the protein recognition site to determine nucleotides important for protein binding. The system was developed with the human protein U1A binding to RNAs derived from U1 snRNA but can also detect RNA-protein interactions in total RNA backgrounds. We propose that this strategy, in combination with emerging coded bead systems, can identify RNAs and RNA sequences important for interacting with RNA-binding proteins on genomic scales. ***Molecular & Cellular Proteomics 1:922–929, 2002.***

RNA-protein interactions are a central component of post-transcriptional regulation at multiple levels including RNA processing, transport, and translation. The sequenced human genome reveals hundreds of potential RNA-binding proteins (1). A critical step toward understanding the function of RNA-binding proteins is to identify and determine how they interact with their target RNAs.

Several strategies have been developed to identify RNA-protein interactions. Genetic approaches include three-hybrid screens (2), phage display (3), and TRAP (translational repression assay procedure) (4) to identify proteins that bind a specific RNA. However, these strategies are generally not applicable to larger RNAs and not suitable for determining binding constants. SELEX (systematic evolution of ligands by exponential enrichment) can identify high affinity RNA sequences that may or may not reflect the biologically relevant recognition site (5).

Recently immunoprecipitation has been combined with microarray analysis to identify candidate RNAs bound to proteins (6, 7). This approach is very promising for inspection of RNA-protein interactions on a genome-wide scale. However, it relies on the ability to preserve stable interactions during immunoprecipitation; many potentially weak interactions may be lost. In addition, extensive motif searching together with additional experimentation may be required to validate the biological significance of any interactors (8).

Recent advances in bead coding technologies to create high complexity platforms are leading to the development of new approaches for high throughput screening studies that could be amenable to the study of RNA-protein interactions (9–12). In principle, nucleic acid hybridization on microbeads offers a number of advantages over DNA chips including shorter hybridization times and better control of binding conditions (13). Therefore, we have developed a new equilibrium binding method on microbeads for elucidating the recognition site of an RNA-binding protein on its cognate RNA. The approach uses oligonucleotide-coupled microbeads and fluorescently labeled RNAs and proteins to monitor RNA-protein binding by flow cytometry. To develop the system for screening RNA-protein interactions, we demonstrate how this approach can be used to identify and characterize the interaction between the spliceosomal protein U1A and a hairpin derived from U1 snRNA as well as detect specific RNAs from total RNA populations.

EXPERIMENTAL PROCEDURES

Plasmids—The U1A test transcript was constructed by annealing overlapping oligos and ligating the annealed product into pDP19 (Ambion) to create plasmid pPS2702. The oligo sequences are: AATTCTTTATCTTCAAAGTTGTCTGTCCAAGATTTGGACTTGTCCGGAGTGCAATGGACG, AAGGACAAGCGTGTCTTCATCAGAGTTGACTTCACTCGAG, GACAAGTCCAAATCTTGGACAGACAACTTTGAAGATAAAG, and GATCCTCGAGTGAAGTCAACTCTGATGAAGACACGCTTGTCCTTCGTCCATTGCACTCCG.

U1A-green fluorescent protein (GFP)[1] was PCR-amplified from pPS2035 and ligated into prSETB (Invitrogen) to create pPS2699. The 96A→G U1A point mutant was constructed by using Stratagene's QuikChange system to create pPS2703. 77C→G was constructed by

From the Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School and The Dana-Farber Cancer Institute, Boston, Massachusetts 02115

Received, October 1, 2002, and in revised form, November 27, 2002

Published, MCP Papers in Press, December 1, 2002, DOI 10.1074/mcp.T200010-MCP200

[1] The abbreviation used is: GFP, green fluorescent protein.

ligating annealed oligonucleotides into pDP19 as described above. All constructs were verified by sequencing.

Transcription and RNA Preparations—PPS2702 and pPS2703 were linearized with *Eco*RI and subsequently transcribed with Ambion's T3 polymerase Maxiscript kit. Labeling with ^{32}P verified a product of the expected size, and subsequent transcription reactions were purified by G-50 spin columns or phenol extractions followed by multiple ethanol precipitations. Texas Red-5-UTP (Molecular Probes) was incorporated during transcription and purified with G-50 spin columns. Total RNA from HeLa cells was prepared by the TRIzol method with high salt precipitations to reduce background GFP fluorescence. Yeast RNA was isolated by a hot phenol method. RNA concentrations were determined by UV spectrometry.

U1A-GFP Purification—Cells were grown to 0.5 optical density before induction with 1 mM isopropyl-1-thio-β-D-galactopyranoside for 4 h. Cells were resuspended in 20 mM HEPES, 10 mM KCl, 0.1% IGEPAL and lysed with lysozyme followed by sonication. After centrifugation, lysate was applied to nickel columns, washed extensively, and eluted with imidazole. Green fractions were pooled and dialyzed into 10 mM HEPES, pH 7.6, 10 mM KCl, 0.1% IGEPAL. To remove the histidine tag, 1.25 units/μl enterokinase were added and incubated for >48 h at 25 °C. Enterokinase was removed with EKaway resin (Invitrogen). U1A-GFP was dialyzed into storage buffer (20 mM HEPES, 20 mM KCl, 0.1% IGEPAL, 10% glycerol). Concentrations were determined by comparing U1A-GFP to bovine serum albumin on Coomassie gels and by the Bio-Rad protein assay. Protein stored at −80 °C bound RNA similarly to fresh preparations (data not shown).

Bead Preparation—Before coupling, Dynal 2.8-μm magnetic streptavidin beads (M-280) were vortexed and/or sonicated to reduce aggregation. Similar to Dynal's recommended protocols, oligonucleotides were attached to beads with 1 nmol of oligonucleotide/9 × 10^6 beads/30 μl. Incubations longer than 5 h were required to reach maximum oligonucleotide density (data not shown). Similar procedures were used for the Spherotech 5.7-μm magnetic streptavidin beads. Oligonucleotides were synthesized with a 12-carbon spacer and 5′ biotin from two different sources: Dana-Farber Cancer Institute Molecular Biology Core Facilities and Integrated DNA Technologies, Inc. Oligonucleotides from each source behaved similarly.

Bead Binding Assays—Binding was performed in 20 mM HEPES, pH 7.5, 300 mM KCl, 0.1% IGEPAL, 10 ng/μl tRNA, 0.04 units/μl superase-IN (Ambion), and 20 ng/μl bovine serum albumin unless otherwise indicated. RNA was heated to 95 °C for 1 min and cooled on ice before being mixed with U1A-GFP for at least 20 min at room temperature before addition of 1 × 10^5 oligonucleotide-coupled beads. Binding reactions were incubated at 35 °C for at least 14 h with constant rotation unless otherwise indicated. Shorter incubations (<6 h) gave lower fluorescence.

Flow Cytometry and Data Analysis—A BD Biosciences Vantage was used to sort beads with both GFP and Texas Red signals. A FACScan was used to monitor GFP alone. Typically 5,000–10,000 beads were counted, and the peak channel, which indicates the maximum height of the bead population, is used to estimate the fluorescence intensity. To determine the percentage of the population shifted to higher fluorescence, cut-offs were set relative to the background fluorescence. Beads with fluorescence above the cut-off are counted in the shifted population. For the binding curves, pro Fit (Quantum Soft) was used to fit the fluorescence intensities to a Langmuir isotherm.

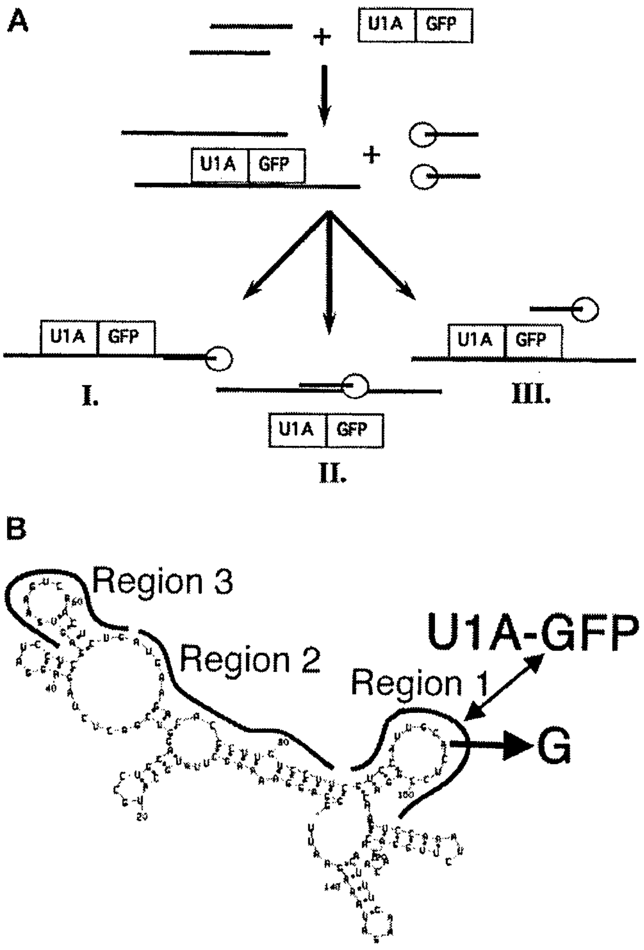


FIG. 1. **Experimental design.** *A*, schematic of the experiment. A U1A-GFP-RNA complex is formed and subsequently challenged with oligonucleotide beads. After reaching equilibrium, RNA and protein fluorescence on each microbead are determined by flow cytometry. The experiment can be performed with or without fluorescently labeled RNA. Three scenarios are possible. *I*, GFP signal is observed indicating the bead is coupled to an oligo complementary to the RNA target but distant from the U1A stem loop recognition site. *II*, no GFP signal is observed, but the oligo is hybridizing to the RNA. With labeled RNA, the RNA-oligo hybridization is detected. These oligos may be complementary to the U1A binding site. *III*, beads with neither GFP nor RNA fluorescence are observed, suggesting that these oligos do not hybridize to the RNA. These sequences may be noncomplementary to the RNA. *B*, predicted secondary structure of U1A RNA constructed for these studies as determined by mFOLD (21). U1A-GFP binds to the hairpin derived from U1 snRNA as indicated. An A to G mutation (96A→G) reduces binding 1000-fold. *Region 1* is complementary to the binding site. Oligonucleotides complementary to other regions of the RNA, distant from the binding site, are also indicated. These sequences are predicted to hybridize to the RNA while U1A-GFP is binding, allowing the U1A RNA-protein interaction to be observed.

RESULTS

The experimental design for analysis of protein-RNA interactions uses oligonucleotides coupled to microbeads to probe RNA-protein interactions and is outlined in Fig. 1*A*. To carry out the analysis, a protein-RNA complex is first formed and then incubated with beads to which oligonucleotides complementary to the target RNA have been coupled. In the experiments described here, RNA is labeled with Texas Red, and the protein is a GFP fusion. After reaching equilibrium, the

TABLE I

Oligonucleotides used in this study

Mismatch nucleotides are in lowercase.

Name	Sequence
	Oligonucleotides complementary to U1A binding site
1.20[a]	TTGTCCGGAGTGCAATGGAC
1.17	GTCCGGAGTGCAATGGA
1.15	GTCCGGAGTGCAATG
	Oligonucleotides complementary to regions distant from U1A binding site
2.20	AAGGACAAGCGTGTCTTCAT
2.17	GGACAAGCGTGTCTTCA
2.15	GGACAAGCGTGTCTT
3.17	TCAGAGTTGACTTCACT
3.15	TCAGAGTTGACTTCA
	Mismatch oligonucleotides
2C.20[b]	AAGGACAAcCGTGTCTTCAT
2C.17	GGACAAcCGTGTCTTCA
2C.15	GGACAAcCGTGTCTT
3C.17	TCAGAcTTGACTTCACT
3C.15	TCAGAcTTGACTTCACT
	U1 snRNA oligonucleotides
4.17	CCCTGCCAGGTAAGTAT
4G.17	CCCTGCgAGGTAAGTAT

[a] The first number indicates the sequence that is being targeted, while the second number indicates the oligonucleotide length, *e.g.* 1.20 is an oligonucleotide complementary to region 1 with 20 nucleotides.

[b] The letter indicates a point mutation in the oligonucleotide disrupting hybridization to the RNA.

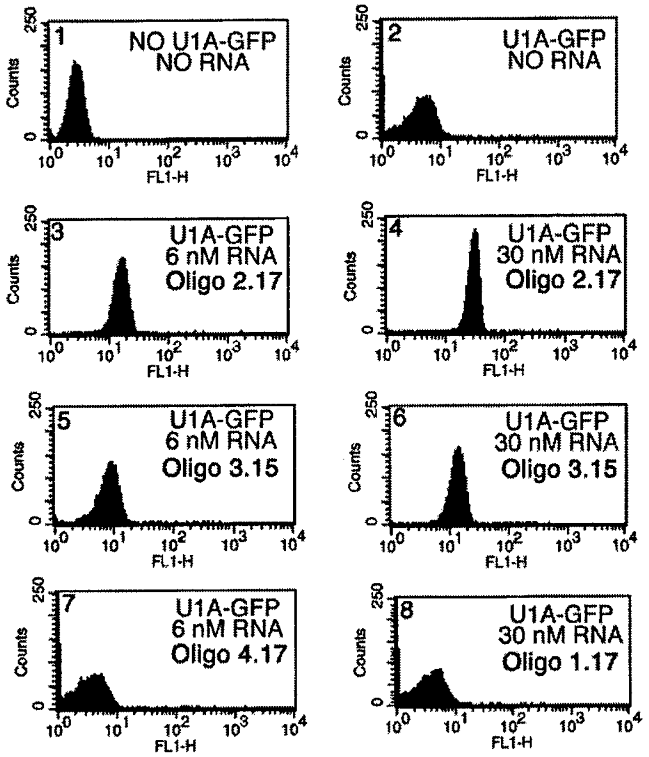


FIG. 2. **Fluorescence is RNA-dependent.** Histograms show the number of beads at different GFP fluorescence intensities. In the presence of RNA and U1A-GFP, the fluorescence intensity of the bead population increases, and a more homogeneous bead population is observed (compare *panels 2*, *3*, and *4*). Oligonucleotides targeting the binding site or not complementary to the RNA show nonspecific binding (compare *panels 2*, *7*, and *8*).

beads are analyzed in a flow cytometer for protein and RNA fluorescence. Beads are sorted into different categories as illustrated in Fig. 1*A*. 1) Beads with both GFP and RNA signals represent the RNA-protein interaction. The oligonucleotides on the beads hybridize to the RNA without interfering with protein binding. 2) RNA signal alone indicates RNA hybridization with no protein binding. The oligonucleotides on these beads may be competing with the protein to bind the same RNA sequences. 3) Some beads will have no detectable fluorescent signal. These oligonucleotide-coupled beads contain sequences that cannot hybridize to the RNA including those that are noncomplementary or contain point mutations. To quantitate the data, the mean fluorescence and/or the percentage of beads in the different categories is determined.

Development of the Bead Binding Assay—The system was developed with the human splicing protein U1A binding to the stem loop 2 derived from U1 snRNA. A U1A-GFP fusion protein including the first 94 amino acids of the RNA recognition motif was expressed with an N-terminal His_6 tag and a C-terminal GFP. The histidine tag was proteolytically cleaved to generate functional U1A-GFP. A 145-nucleotide RNA encoding random sequence and including the specific U1A hairpin was designed. The predicted secondary structure is shown in Fig. 1*B*. Gel shift mobility experiments confirmed that U1A-GFP binds the U1A RNA with a dissociation constant of ~35 nM in 150 mM KCl at 25 °C (data not shown), similar to that reported for the same 94-amino acid fragment in the absence of the GFP (14). Oligonucleotides complementary to different regions of the U1A RNA were designed as illustrated in Fig. 1*B* and listed in Table I. Oligonucleotides were synthesized with a 5′ biotin label and attached to streptavidin beads. Reproducible coupling conditions were devised to ensure similar oligonucleotide concentrations per bead as determined with ^{32}P-labeled oligos (data not shown). The oligonucleotide concentrations used in the binding assay are estimated to be between 1–2 nM. For all experiments, Dynal 2.8-μm diameter streptavidin beads were used unless otherwise indicated.

After the RNA-protein complex is formed, oligo-coupled beads are added. To reach equilibrium, incubations at 25 °C or 35 °C for longer than 6 h were necessary (data not shown) and typical incubations were at least 14 h. After reaching equilibrium, GFP fluorescence on individual beads was assessed in a flow cytometer. The RNA dependence and specificity of the binding reactions were assessed as follows to ascertain the validity of the approach.

The observed GFP fluorescence on the beads is RNA-dependent as illustrated in Fig. 2*A*. Background U1A-GFP binding to the beads is low, and the peaks are broad indicative of a relatively heterogeneous population (Fig. 2, *panel 2*). The observed GFP fluorescence intensity increases with higher U1A RNA concentrations, and the population is more homogenous as indicated by the narrower peak width (Fig. 2, *panels 3* and *4*). The observed signals were reproducible with different RNA and protein preparations.

The assay accurately distinguishes the U1A RNA binding site. Oligonucleotides complementary to the binding site compete with U1A-GFP for the same RNA sequence and thereby reduce observed GFP fluorescence. Only nonspecific background GFP signal is observed for oligos complementary to the binding site (Fig. 2, compare *panels 2* and *8*). On the other hand, oligos complementary to sequences not part of the U1A binding site show significant GFP signal (oligos 2.17 and 3.15, Fig. 2*A*, *panels 3–6*). These oligonucleotides are hybridizing to the RNA without interfering with U1A-GFP binding thereby allowing the observation of the RNA-protein interaction. As a control, oligonucleotides not complementary to the RNA show background nonspecific signal (Fig. 2, compare *panels 2* and *7*).

Decreasing the oligonucleotide length lowers the GFP fluorescence intensity. 20-mers, 17-mers, and 15-mers all yield significant fluorescence, while 10-mers complementary to the same region do not (data not shown). Lower fluorescence is consistently observed for 15-mers, such as oligo 3.15, compared with 17-mers, such as oligo 2.17 (Fig. 2, compare *panels 3* and *4* to *panels 5* and *6*). These observations are not limited to a particular region of the RNA or sequence.

Discrimination of Oligonucleotide Mismatches—The bead assay discriminates between oligonucleotides that contain mismatches under conditions that preserve RNA-protein interactions. When mismatches in the middle of the complementary sequence are introduced, the oligonucleotide yields significantly lower GFP fluorescence. Mismatch discrimination is not unique to a particular sequence as oligonucleotides complementary to distinct regions show a significant difference in GFP fluorescence (Fig. 3*a*, compare oligos 2.17 and 2C.17 and oligos 3.17 and 3C.17). Interestingly, unlike 15-mers and 17-mers, 20-mers do not discriminate mismatches as well (Fig. 3*a*, compare oligos 2.20 and 2C.20 with oligos 2.17 and 2C.17). Also mismatches at the first or second position of either end of the oligonucleotide are not discriminated as well as those in the middle of the sequence (data not shown).

To verify the observed oligonucleotide mismatch discrimination, a compensatory mutation in the RNA was constructed. Binding reactions were prepared with the two different RNAs, and beads coupled to either oligo 2.17 or 2C.17 were added. As observed previously for the wild-type U1A RNA, oligo 2.17 shows significant GFP fluorescence, while oligo 2C.17 does not (Fig. 3*b*). However, the compensatory mutation in the U1A RNA, 77C→G, creates a mismatch for oligo 2.17 and disrupts the hybridization thereby reducing the observed GFP fluorescence. Meanwhile, significant GFP fluorescence is observed for oligo 2C.17, which is complementary to 77C→G RNA (Fig. 3*b*). Compensatory RNA mutations and subsequent U1A binding have also been performed with oligos 3.17 and 3C.17 (data not shown). These "rescue" experiments further verify the observed point mutant hybridization discrimination.

The observed mismatch discrimination is enhanced by measuring hybridization through U1A-GFP binding. In the absence of U1A-GFP, poor discrimination between oligonucleotide mismatches is observed (Fig. 3*c*, compare *panels D* and *E*). This is consistent with reports of poor hybridization behavior of short oligonucleotide sequence (15, 16). However, in the presence of U1A-GFP, the same oligo 2.17 beads show both a higher GFP and Texas Red fluorescence intensity, while two different mismatches of oligo 2.17 show significantly reduced GFP fluorescence (Fig. 3*c*). Thus, in physiological conditions, mismatch discrimination is observed by monitoring hybridization through an RNA-protein interaction.

Detection of Specific RNA-Protein Interactions—To determine whether the observed fluorescence is accurately reflecting the U1A RNA-protein interaction, an A to G point mutation in the U1A loop (96A→G) known to disrupt binding was tested (14). This mutation severely reduces the observed GFP signal as shown in Fig. 4*A*. To quantitate Texas Red and GFP fluorescence, the percentage of the bead population shifted beyond the signals observed for background binding is determined. The quadrants shown in *panels A–C* of Fig. 3*c* determine the cut-offs to define the bead populations with different combinations of Texas Red and GFP fluorescence. Oligos 2.17 and 3.17 show U1A-GFP signal with wild-type U1A RNA, while only background fluorescence is observed with 96A→G. Importantly both oligonucleotides are hybridizing to the RNA as indicated by significant Texas Red fluorescence suggesting that the lower GFP signal is due to disruption of the protein-RNA complex and not reduced hybridization.

As predicted, oligo 1.17, complementary to the U1A binding site, does not show any significant U1A-GFP binding, similar to the experiments described above. Importantly oligo 1.17 is hybridizing to the RNA at levels similar to other oligonucleotides where GFP fluorescence is observed. This suggests that the low observed GFP fluorescence is due to disruption of the RNA-protein complex and not poor hybridization. These data demonstrate the ability of this system to define sequences important for protein recognition on the RNA by footprinting.

To further demonstrate that the observed GFP fluorescence accurately reflects the RNA-protein interaction, the affinity of the U1A complex was measured (Fig. 4*B*). A 75 nM dissociation constant in 300 mM KCl at 25 °C is estimated by curve fitting to a Langmuir isotherm, consistent with published data (14). Meanwhile, the 96A→G point mutant shows no significant binding under the same conditions, consistent with its ~1000-fold weaker affinity for this U1A construct (14). Higher nonspecific U1A-GFP binding to the beads causes broader bead population distributions and is probably responsible for the larger error bars observed at higher protein concentrations.

The observed GFP fluorescence is also sensitive to the salt

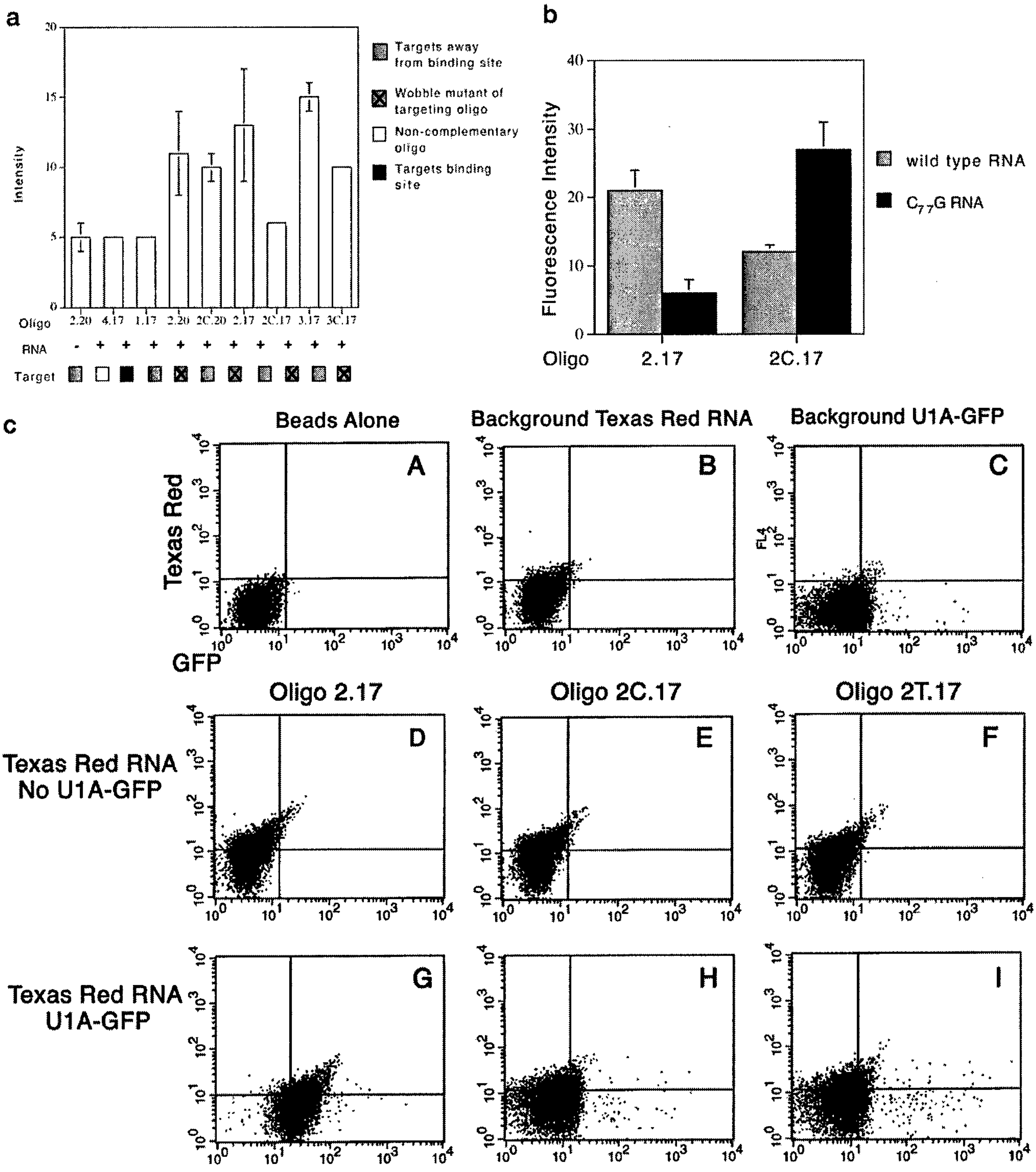


FIG. 3. **Mismatches are discriminated when monitoring hybridization indirectly through U1A-GFP binding.** *a*, the bar graph shows the mean fluorescence of observed GFP signal of bead populations. Fluorescence intensity is in arbitrary units. Triplicate data are averaged, and the *error bars* represent standard deviations. When the standard deviation is less than 1, no error bar is shown. Binding reactions include 75 nM U1A-GFP and 6 nM U1A test RNA. *b*, compensatory RNA mutations restore U1A-GFP binding for an oligonucleotide point mutant. Oligo 2.17 hybridizes to the RNA and shows U1A-GFP binding, while 2C.17 does not. 77C→G, which is an exact match to oligo 2C.17, disrupts hybridization to oligo 2.17 and allows hybridization to oligo 2C.17. Binding reactions include 10 nM U1A RNA and 100 nM U1A-GFP. *Error bars* represent standard deviations of duplicates. *c*, plots show Texas Red fluorescence on the *y* axis and GFP fluorescence on the *x* axis. *Panels A–C* show the nonspecific binding of Texas Red RNA and U1A-GFP to the beads. *Panels D–F* illustrate the poor discrimination of 17-mer oligos

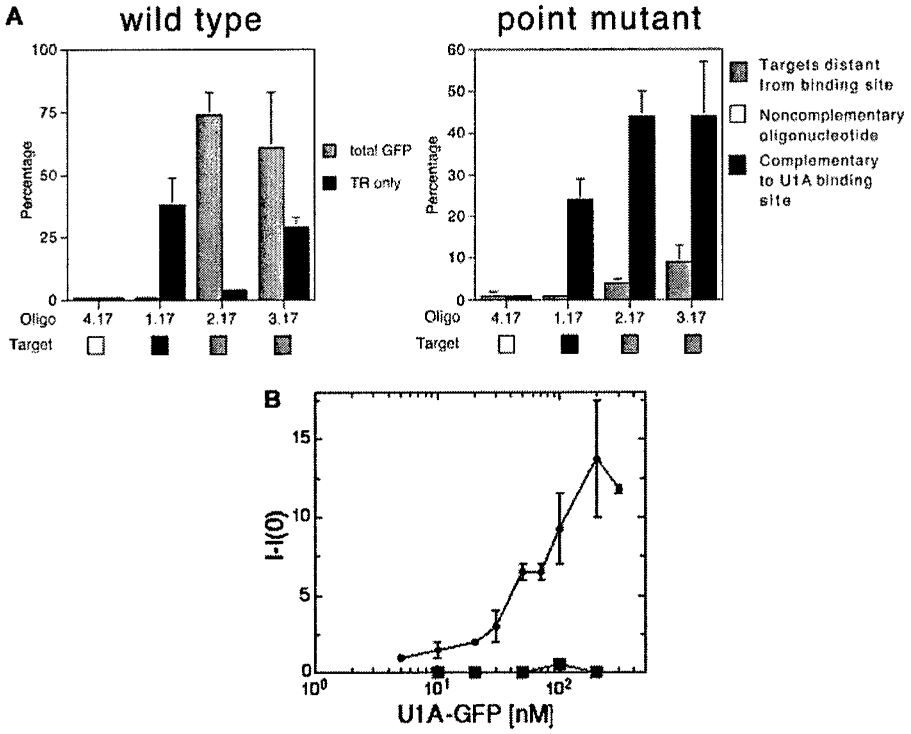


FIG. 4. **Bead assay accurately reflects the RNA-protein interaction.** *A*, percentage of beads shifted above background are tabulated into different categories: beads with Texas Red fluorescence only and beads with GFP fluorescence. U1A-GFP does not bind the 96A→G point mutant RNA, but oligonucleotides still significantly hybridize to the RNA as indicated by the Texas Red fluorescence. *Error bars* represent standard deviations of triplicate experiments. *B*, titration of U1A-GFP into 3 nM RNA generates binding curves. *Squares* represent the 96A→G point mutant, and *circles* represent the wild-type U1A RNA. Background fluorescence (I_0) in the absence of RNA is subtracted for each protein concentration. Binding curves were performed in duplicate and averaged. *Error bars* represent standard deviations. Curve fitting suggests U1A-GFP is binding with a dissociation constant of ~75 nM in 300 mM KCl at 25 °C.

concentration. At higher KCl concentrations, the GFP fluorescence intensity increases by ~25% at each protein concentration. However, the dissociation constant shifts from ~35 to ~75 nM. These observations reflect stronger hybridization and weaker U1A interaction at higher salt concentrations. In sum, these data demonstrate the specificity and affinity of the U1A RNA-protein interaction on beads.

Specific Binding in Mixed Populations—For screening RNA-protein interactions with the bead assay, RNAs will have to be identified from complex mixtures of RNAs. To determine whether total RNA can compete with U1A-GFP binding, yeast RNA was added to the binding reactions. Human U1A does not specifically bind any yeast RNA (17, 18). Only in the presence of the U1A test RNA is GFP fluorescence detectable with oligo 2.17 as shown in Fig. 5*A*. Oligonucleotides not complementary to the U1A RNA such as oligo 4.17 show fluorescence equivalent to background. This suggests that even in contexts where there might be high nonspecific binding, specific binding is observable. These experiments were performed with larger 5.7-μm diameter beads because the sensitivity is higher (data not shown). The nonspecific binding on these larger beads is also higher as more GFP fluorescence is observed. However, specific GFP fluorescence is observed at lower concentrations of U1A-GFP compared with the smaller 2.8-μm microbeads.

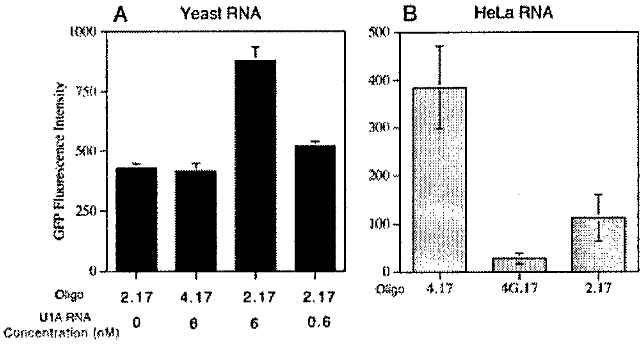


FIG. 5. **U1A-GFP binding to the U1A test RNA and to U1 snRNA is detectable in total RNA backgrounds.** *A*, oligo 2.17 but not oligo 4.17 detects U1A-GFP signal in 0.1 μg/μl yeast total RNA and 25 nM U1A-GFP. Fluorescence intensities are higher with larger 5.7-μm beads. Data are in triplicate with standard deviation *error bars*. *B*, conversely oligo 4.17 shows GFP fluorescence in HeLa total RNA. U1A-GFP is binding to snRNA in the total HeLa RNA. Oligo 4.17 also shows higher GFP fluorescence than a mismatch, oligo 4G.17, or a noncomplementary sequence, oligo 2.17. Reactions include 0.16 μg/μl HeLa total RNA (40 μg total) and 50 nM U1A-GFP. *Error bars* represent standard deviations of triplicates. Background signals were subtracted for this graph.

To determine whether RNAs isolated from RNA preparations can be identified with microbeads, total RNA was isolated from HeLa cells and mixed with U1A-GFP. In HeLa RNA,

binding to the RNA in the absence of U1A-GFP. Oligonucleotide mismatches do not significantly affect the observed Texas Red signal. Only the wild-type oligonucleotide shown in *panel G* gives significant GFP fluorescence with 100 nM U1A-GFP and 30 nM U1A RNA, while the mismatches, in *panels H* and *I*, show some Texas Red signal but no significant GFP fluorescence.

U1A binds to U1 snRNA and its own mRNA. An oligonucleotide complementary to snRNA, oligo 4.17, shows higher GFP fluorescence compared with a mismatch, oligo 4G.17, and a noncomplementary oligonucleotide, oligo 2.17, as illustrated in Fig. 5*B*. The U1 snRNA concentration is ~1–5 nM in these experiments. The observed GFP fluorescence is HeLa RNA-dependent as the signal varies with HeLa RNA concentration but remains unchanged with increasing equivalent concentrations of yeast RNA. In sum, these data demonstrate that a specific RNA can be identified with microbeads from total RNA.

An additional requirement for screening with oligonucleotide bead libraries is the ability to detect a small percentage of oligonucleotide beads from a large background of beads that do not bind. Initial experiments diluting the oligonucleotide beads 100-fold suggest that at even low oligonucleotide concentrations the RNA-protein complex can be detected. Further experiments diluting oligo 2.17 in a large excess of oligo 4.17 demonstrated that two populations of beads could be differentiated by GFP fluorescence (data not shown). These data also demonstrate that the observed shifts are sufficient to identify an RNA-protein complex.

DISCUSSION

We describe a system to monitor RNA-protein interactions in solution with microbeads using flow cytometry. We demonstrate the versatility of the approach for 1) discriminating between mismatches in the oligonucleotides, 2) mapping protein recognition sites on RNA, 3) differentiating specific and nonspecific binding RNAs, and 4) detecting specific RNAs in complex mixtures. Importantly, specific binding can be detected in high nonspecific RNA backgrounds, and the system can discriminate a nonspecific binding point mutant at a variety of protein and RNA concentrations. Because flow cytometry is used to monitor the fluorescence on distinct oligonucleotide-coupled microbeads, the system is amenable to high throughput, genomic scale screening of RNA-protein interactions.

With the U1A interaction, we have determined the fundamental requirements for using this microbead system for screening RNA-protein interactions. The approximate 2-fold changes in observed GFP fluorescence are sufficient to distinguish specific protein binding to RNA from background. Future versions of the system may have increased sensitivity and dynamic range by using brighter fluorophores and microbeads with higher oligonucleotide densities. Furthermore, because the system is monitoring the protein binding to a distribution of thousands of microbeads, the fluorescence shifts are more significant than just monitoring the bulk signal. Methods to analyze the distributions more quantitatively are being developed. The microbead system also allows for the measurement of relative affinities of a protein for its cognate RNA thereby distinguishing specific and nonspecific binding during screening and thereby reducing false positives. Two approaches are possible. After initial screening at a particular protein concentration, binding experiments will help distinguish specific and nonspecific binding candidates. Alternatively, screening at different protein concentrations could be performed to determine relatively strong and weak binding interactions.

The microbead-based system described here has a number of advantages over other recently developed RNA-protein screening strategies. It is rapid with few time-consuming manipulations required. Also, unlike many *in vivo* strategies, there are no limitations to the size of the RNA or its basic structural features. It is often difficult to monitor the binding of large RNAs directly because they do not migrate well in gel electrophoresis for mobility shift studies. With the microbead system, the only requirement is that oligonucleotides hybridize to regions distant from the binding site. Most mRNAs have multiple regions accessible to hybridization in physiological conditions (19).

With the system presented here, hybridization is monitored indirectly through the protein fluorescence. To observe protein fluorescence, the oligonucleotides need to hybridize to particular RNAs that are also binding the protein. The combination of observed RNA hybridization and protein fluorescence on a microbead indicates that an RNA-protein complex is present. Thus, many RNAs may be hybridizing to the beads, but only when the fluorescently labeled protein is bound with high affinity to one of these RNAs is positive signal observed. This significantly reduces the nonspecific binding that would be observed in identifying possible RNA targets compared with other strategies that isolate all the RNAs bound to beads. In its current form, the system does not reach equilibrium for hours presumably because of slow hybridization at physiological temperatures. Smaller volumes may help reduce the time required to reach equilibrium.

The mismatch discrimination observed with this bead strategy may allow it to be adopted for single nucleotide polymorphism analysis. Similar to assays such as the invasive cleavage method (20), monitoring hybridization indirectly provides a sensitivity enhancement to observe the subtle effects of mismatches on hybridization. The enhanced mismatch discrimination observed through the protein interaction may be particular to the U1A system. Future studies of other RNA-binding proteins will determine the generality of the observed mismatch discrimination.

For genomic screening, proteins bound to RNA could be challenged with large oligonucleotide coded bead libraries. The coded beads would be sorted in a flow cytometer while monitoring RNA and protein fluorescence to determine which sequences are hybridizing to the RNA while preserving the RNA-protein complex. This information can then be compared with sequenced genomes to determine which RNAs are binding and which sequences may be important for the interaction. Various coding strategies are currently being developed that do not require decoding or very large beads (9–12).

In sum, we have developed a microbead-based system to determine which RNAs may be binding a particular protein as well as which RNA sequences may be important for the RNA-protein interaction. Many applications of the assay are possible including binding in cell extracts, single nucleotide polymorphism analysis, and monitoring the effects of small molecules on RNA-protein complexes. Perhaps the most inviting aspect of this system is to use large coded oligonucleotide bead libraries to probe RNA-protein interactions on genomic scales.

Acknowledgments—We thank Kathy Hall for U1A plasmids, Vicki Losick for constructing 96A→G, and the staff at the Dana-Farber Cancer Institute flow cytometry facility. We thank M. Damelin, C. Stern, P. Zarrinkar, and D. Treiber for critical reading of the manuscript and Jeff Way and Matt Trau for stimulating discussions.

* A portion of this work was supported by grants from the National Institutes of Health and the Department of Defense (to P. A. S.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ Supported by the Claudia Barr Young Investigator program.

§ To whom correspondence should be addressed: Dana-Farber Cancer Inst., SM 922, 1 Jimmy Fund Way, Boston, MA 02115. Tel.: 617-632-5105; Fax: 617-632-5103; E-mail: alex_brodsky@dfci.harvard.edu.

REFERENCES

1. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921
2. Zhang, B., Kraemer, B., SenGupta, D., Fields, S. & Wickens, M. (2000) Yeast three-hybrid system to detect and analyze RNA-protein interactions. *Methods Enzymol.* **318,** 399–419
3. Danner, S. & Belasco, J. G. (2001) T7 phage display: a novel genetic selection system for cloning RNA-binding proteins from cDNA libraries. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 12954–12959
4. Paraskeva, E. & Hentze, M. W. (2000) Translational repression assay procedure: a method to study RNA-protein interactions in yeast. *Methods Enzymol.* **318,** 374–384
5. Andrews, L. G. & Keene, J. D. (1999) Identification of specific protein-RNA target sites using libraries of natural sequences. *Methods Mol. Biol.* **118,** 233–244
6. Tenenbaum, S. A., Carson, C. C., Lager, P. J. & Keene, J. D. (2000) Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 14085–14090
7. Brown, V., Jin, P., Ceman, S., Darnell, J. C., O'Donnell, W. T., Tenenbaum, S. A., Jin, X., Feng, Y., Wilkinson, K. D., Keene, J. D., Darnell, R. B. & Warren, S. T. (2001) Microarray identification of FMRP-associated brain mRNAs and altered mRNA translational profiles in fragile X syndrome. *Cell* **107,** 477–487
8. Darnell, J. C., Jensen, K. B., Jin, P., Brown, V., Warren, S. T. & Darnell, R. B. (2001) Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function. *Cell* **107,** 489–499
9. Battersby, B. J., Lawrie, G. A. & Trau, M. (2001) Optical encoding of microbeads for gene screening: alternatives to microarrays. *Drug Discov. Today* (*HTS Suppl.*) **6,** S19
10. Chan, W. C., Maxwell, D. J., Gao, X., Bailey, R. E., Han, M. & Nie, S. (2002) Luminescent quantum dots for multiplexed biological detection and imaging. *Curr. Opin. Biotechnol.* **13,** 40–46
11. Cao, Y. C., Jin, R. & Mirkin, C. A. (2002) Nanoparticles with Raman spectroscopic fingerprints for DNA and RNA detection. *Science* **297,** 1536–1540
12. Cunin, F. *et al.* (2002) Biomolecular screening with encoded porous-silicon photonic crystals. *Nat. Mater.* **1,** 39–41
13. Wilkins Stevens, P., Henry, M. R. & Kelso, D. M. (1999) DNA hybridization on microparticles: determining capture-probe density and equilibrium dissociation constants. *Nucleic Acids Res.* **27,** 1719–1727
14. Zeng, Q. & Hall, K. B. (1997) Contribution of the C-terminal tail of U1A RBD1 to RNA recognition and protein stability. *RNA* **3,** 303–314
15. Guo, Z., Guilfoyle, R. A., Thiel, A. J., Wang, R. & Smith, L. M. (1994) Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports. *Nucleic Acids Res.* **22,** 5456–5465
16. Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. & Brown, E. L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14,** 1675–1680
17. Tang, J. & Rosbash, M. (1996) Characterization of yeast U1 snRNP A protein: identification of the N-terminal RNA binding domain (RBD) binding site and evidence that the C-terminal RBD functions in splicing. *RNA* **2,** 1058–1070
18. Brodsky, A. S. & Silver, P. A. (2000) Pre-mRNA processing factors are required for nuclear export. *RNA* **6,** 1737–1749
19. Ho, S. P., Bao, Y., Lesher, T., Malhotra, R., Ma, L. Y., Fluharty, S. J. & Sakai, R. R. (1998) Mapping of RNA accessible sites for antisense experiments with oligonucleotide libraries. *Nat. Biotechnol.* **16,** 59–63
20. Wilkins Stevens, P., Hall, J. G., Lyamichev, V., Neri, B. P., Lu, M., Wang, L., Smith, L. M. & Kelso, D. M. (2001) Analysis of single nucleotide polymorphisms with solid phase invasive cleavage reactions. *Nucleic Acids Res.* **29,** E77
21. Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288,** 911–940